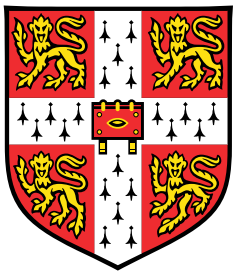
Analyses of RNA dynamics in *Mus musculus*



Wayo Matsushima

Department of Genetics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

St John's College

April 2020

この博士論文を敬愛する両親に捧ぐ。

I dedicate this thesis to my loving parents.

Declaration

I hereby declare that this thesis is the result of my own work, and all the collaborative works are declared in the Acknowledgements and the text.

The contents of this dissertation are original and not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

This dissertation contains fewer than 60,000 words, exclusive of tables, footnotes, bibliography, and appendices.

Wayo Matsushima
April 2020

Acknowledgements

I would like to thank Professor Eric Miska for being my PhD supervisor. He has always been supportive of my PhD, and his thirst for exciting science has been motivational.

Miska Lab members have also been supportive throughout my PhD. They contribute to lab meetings, and I learned a lot through critical and constructive discussions with the members. The lab's friendly and culturally diverse environment has made me feel at home and enabled me to overcome the toughest times of my PhD. I would like to particularly mention Katharina Gapp, who also worked on mice and was a great mentor for my mouse-related projects. She has also taken care of the most time-intensive administrative procedures for mouse experiments, which made my PhD projects possible. Also, Tomás Di Domenico was a exceptional mentor in bioinformatics and trained me to perform computational analyses independently. Navin Ramakrishna took his time to give me constructive comments on a few chapters of my thesis. I also thank Marc Ridyard and Miranda Landgraf for their excellent lab management.

My PhD project utilised a novel method developed in the Stefan Ameres Lab at the Institute of Molecular Biotechnology (IMBA) in Austria. I would like to thank Stefan Ameres and his colleagues, especially Veronika Herzog and Brian Reichholf, for their tremendous experimental support in establishing my method. Also, I thank Tobias Neumann from the Johannes Zuber Lab at the Institute of Molecular Pathology (IMP) for devoting a lot of his time for me to establish a bioinformatic pipeline to analyse the high-throughput sequencing data. This experience has taught me the beauty of international scientific collaboration.

I have also received substantial experimental supports at Sanger Institute and Gurdon Institute. Research Support Facility (RSF) staff maintained mice and assisted me with many of the experiments. Evelyn Grau has collected zygotes from mice and cultured them for my experiments. High-throughput sequencing libraries were quality-checked and the runs were performed thanks to valuable support from Kay Harnish.

I would also like to thank my academic advisor, David Summers. He has kindly arranged frequent supervisions and ensured my PhD was on-track.

My PhD at Cambridge would have never been possible without the people I met prior to even setting foot in England. Professor Keiko Nakayama at Tohoku University kindly gave

me the opportunity to undertake a research project in her lab and taught me the fundamentals and joys of research. Tom Misteli from the National Cancer Institute (NCI) accepted me as a summer student and supported every aspect of my application to graduate programs abroad.

I have been financially supported by two generous funding bodies: The Nakajima Foundation and St John's College Benefactors' Scholarship. They not only covered necessary costs to study in Cambridge, but also supported my conference travels abroad.

I would like to show my gratitude to Marii Konishi for being with me and taking time to proofread my thesis. I am fully responsible for any remaining errors.

Finally, I would like to express my deepest gratitude to my parents and family members, who have unconditionally supported my ambition.

Abstract

Ribonucleic acid (RNA) is a biological molecule that exists in the cell of virtually all kinds of known living organisms. Although its significance in fundamental cellular processes such as protein synthesis has been known for decades, an increasing number of novel species and functions of RNA have been discovered recently, and a complete picture of RNA functions in the cell is still elusive. One of the major challenges in RNA studies is that there had been no efficient method to quantify RNA with spatial and temporal information *in vivo*. In my doctoral studies, I developed a novel RNA sequencing method that metabolically labels RNA in a cell- and time-specific manner in mice and applied this method to solve biological problems that had been difficult to tackle.

First, the robustness of the newly developed *in vivo* RNA labelling method was assessed. To confirm the sensitivity and specificity of RNA labelling, multiple transgenic mice that label RNA in different cell types were generated. By comparing the data obtained from each transgenic strain to previously generated transcriptomic datasets, I confirmed that RNA labelling in a specific cell type was achieved in all the strains analysed. This method would be useful to study cell-type-specific transcriptomics rather than the commonly used laborious and time-intensive cell isolation method often used, and might provide data that closely reflect the native transcriptional state *in vivo*.

Next, using the same RNA labelling method, I tested if there is any RNA that is mobile between different cell types in mice. Intercellular mobility of RNA has been shown in nematodes and plants, but whether there is any RNA that is mobile between different mammalian cells *in vivo* is still unclear. The cell-specific RNA labelling method allowed us to assess the mobility of RNA directly for the first time. Based on previous publications, three different cell types were chosen as potential "donor" cells, and transgenic mice that label RNA in these cells were generated. The donor cell-derived labelled RNA was sought in "recipient" tissues that are not capable of labelling RNA. However, although RNA labelling was achieved in the donor tissues, no labelled RNA was found in the recipient tissues in any of the animals tested. This experiment presents a novel methodology to assess the mobility of RNA in living mice, and the obtained data suggest that only minor intercellular transfer of RNA, at best, is happening in the tested pairs of tissues.

In the final part of my thesis, I applied the metabolic RNA labelling method to study the transcriptional dynamics in the early preimplantation mouse embryos. Unlike conventional RNA sequencing methods that can only quantify RNA abundance in each stage of the embryos, metabolic RNA sequencing can directly interrogate the transcriptional activity of each gene. This method is particularly powerful in studying the transcriptional network in the early preimplantation embryo, where embryo-derived RNA needs to be distinguished from maternally-deposited RNA. By exposing the mouse embryos to a nucleotide analogue in a stage-specific manner, I identified genes that are actively transcribed in the 2-cell embryos. This method would be useful in studying the transcriptional cascade in the early mammalian preimplantation embryos.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

2C	2-cell
4-thio-UMP	4-thiouridine monophosphate
4-thio-UTP	4-thiouridine triphosphate
E. coli	*Escherichia coli*
T. gondii	*Toxoplasma gondii*
A	adenine
APRT	adenine phosphoribosyltransferase
BAT	brown adipose tissue
BrU	5-bromouridine
CA	chicken beta-actin
C	cytosine
CMP	cytidine monophosphate
cpm	counts per million
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DRB	5,6-dichloro-1-β-D-ribofuranosyl-benzimidazol

dsRNA	double-stranded RNA
DTT	dithiothreitol
endo-siRNA	endogenous siRNA
EU	5-ethynyluridine
eWAT	epididymal white adipose tissue
FACS	fluorescence-activated cell sorting
FDR	false discovery rate
FPKM	fragments per kilobase of transcript per million mapped reads
FU	fluorescence unit
GFP	green fluorescent protein
G	guanine
GO	gene ontology
GTFs	general transcription factors
hCG	human chorionic gonadotropin
HGPRT	hypoxanthine-guanine phosphoribosyltransferase
HSV	human sarcoma virus
HTS	high-throughput DNA sequencing
i.p.	intraperitoneal
IAA	iodoacetamide
IMP	inosine monophosphate
KO	knock-out
KSOM	potassium-supplemented simplex optimised medium
LCM	laser capture microdissection
lncRNA	long non-coding RNA

loxP	locus of X-over P1
miRNA	microRNA
mRNA	messenger RNA
MuERV-L	murine endogenous retrovirus with leucine tRNA primer
NMP	nucleoside monophosphate
nt	nucleotide(s)
NTP	nucleoside triphosphate
ORF	open reading frame
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
piRNA	PIWI-interacting RNA
Pol II	RNA polymerase II
pre-miRNA	precursor miRNA
pri-miRNA	primary miRNA
PRPP	phosphoribosyl pyrophosphate
RIN	RNA integrity number
RNA	ribonucleic acid
RNase	ribonuclease
rRNA	ribosomal RNA
RT-qPCR	reverse transcription followed by quantitative polymerase chain reaction
RT	reverse transcription
s^4U	4-thiouridine
scRNA-seq	single-cell RNA-seq
siRNA	small interfering RNA

SNP	single-nucleotide polymorphism
T>C	thymine-to-cytosine
TE	transposable element
TF	transcription factor
tRF	tRNA-derived fragment
tRNA	transfer RNA
UDP	uridine diphosphate
UMI	unique molecular identifier
UMP	uridine monophosphate
UPRT	uracil phosphoribosyltransferase
UTP	urdine triphosphate
UTR	untranslated region
U	uracil
UV-Vis	ultraviolet-visible
WAT	white adipose tissue
WT	wild-type
ZGA	zygotic genome activation

Chapter 1

Introduction

1.1 Ribonucleic acid

Existence of nucleic acid was first reported by a Swiss chemist, Friedrich Miescher, in 1871. It was named "nuclein" because it was discovered in nuclei of leucocytes isolated from the pus on surgical bandages (Dahm, 2005). Later, studies on different tissues from various organisms revealed that there are two kinds of nucleic acids, which were initially called thymus nucleic acid and yeast nucleic acid, based on what these molecules are extracted from. This was, based on our current knowledge, because these tissues have a different ratio of deoxyribonucleic acid (DNA) to ribonucleic acid (RNA).

The contribution of RNA in protein synthesis process was predicted by its existence in the cytoplasm, where protein synthesis also takes place. Also, metabolic labelling experiments revealed a correlation between the protein and RNA synthesis rate (Gorman and Halvorson, 1959). Although scientists discovered RNA-rich particles called "microsomal particles", which we now know as ribosome, and speculated its significance in protein synthesis as early as the mid 1950s (Littlefield et al., 1955), understanding how exactly the genetic code is translated through RNA had been challenging. This is partly due to the extensive disproportion of RNA species; while ribosomal RNA (rRNA) and transfer RNA (tRNA) account for 85% and 10% of cellular RNA, respectively, messenger RNA (mRNA) only accounts for 5%. Also, mRNA has a much shorter half-life and is quite unstable compared with the rest of the RNA species. Thus, scientists at the time only focused on the "soluble RNA", or tRNA, and "stable RNA", which is rRNA residing in microsomal particles, and thought that the stable RNA was the template for protein synthesis (Fig. 1.2).

mRNA was first observed as a distinct class of molecules from the other RNA by Al Hershey's group in 1953 and also by Volkin and Astrachan in 1958 (Astrachan and Volkin, 1958; Hershey et al., 1953). Both groups exposed bacteria to radiolabelled nucleic acid

precursors followed by a phage infection, and observed short-lived RNA that was clearly distinct from tRNA and rRNA. However, at this point, its link to protein synthesis remained obscure.

A conceptual result that suggests the existence of a messenger molecule was shown by the experiments performed by Pardee et al. (1959), which is better known as the PaJaMa experiment named after the authors' surnames (Pardee, Jacob, and Monod). This experiment is primarily important in discovering the concept of gene regulation. They used *Escherichia coli* (*E. coli*) with two alleles: β-galactosidase Repressor mutant (i^-) and β-galactosidase mutant (z^-) (Fig. 1.1). They showed that *E. coli* (i^- z^-) started synthesising β-galactosidase after being crossed with wild-type *E. coli* (i^+ z^+) within a minute (Pardee, 2002). This experiment suggested that there is a system that works as "a messenger" from a gene to the host cell protein synthesis system. Three hypotheses were presented to explain the phenomenon (Fig. 1.2) and were elegantly tested in back-to-back papers published in *Nature* in 1961. Two groups showed that the previously observed transient RNA synthesised in response to the phage infection is the messenger that conveys the genetic information for protein synthesis (Brenner et al., 1961; Crick et al., 1961). Brenner et al. exposed bacteria to different radiolabelling agents before and after an infection, and showed that the newly synthesised RNA after infection indeed interacted with ribosomes. Also, no newly synthesised ribosomes were detected, while newly transcribed RNA had the similar base composition to that of the phage DNA. These results collectively suggest that the model III is the most plausible model (Fig. 1.2).

1.2 Regulatory RNA

It was not too long after the discovery of mRNA before people noticed that there were classes of RNA that are not translated into protein. Notably, most of the abundant classes of RNA in a cell are non-protein-coding but play important roles in protein synthesis. For instance, ribosomal RNA (rRNA), together with ribosomal proteins, join amino acids based on the mRNA sequence to synthesise proteins, and, importantly, rRNA itself is the key component that has peptidyl transferase activity to join the amino acids (Noller et al., 1992). Hoagland and Zamecnik successfully identified the soluble RNA, tRNA, as a transporter of amino acids to ribosome (Hoagland et al., 1958). Also, small nuclear RNAs (snRNAs), which function as core components of RNA splicing machinery was also discovered (Lerner et al., 1980; Zhuang and Weiner, 1986). This series of discoveries clearly proved that some RNA species act as functional molecules in essential biological systems without being translated into proteins.

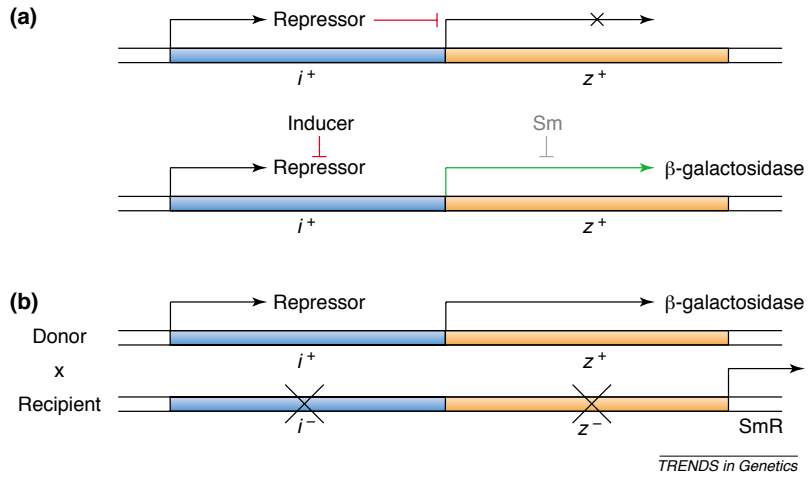


Fig. 1.1 PaJaMa experiment
Schematic of experimental system utilised by Pardee, Jacob, and Monod (taken from Pardee (2002)). (a) In wild-type *E. coli*, Repressor transcribed from *i* gene inhibits the transcription of β-galactosidase from *z* gene. Repressor production is inhibited by Inducer (galactosidase) and *z* starts to synthesise β-galactosidase. (b) In the PaJaMa experiment, genes from wild-type (i^+ z^+) were transferred to mutant bacteria (i^- z^-). Since there is no Repressor present in the recipient bacteria, β-galactosidase was initially synthesised. However, as Repressor accumulates in the cell, β-galactosidase gradually diminished.

A large non-coding RNA class that does not synthesise mRNA but contributes to regulate the abundance of RNA by controlling the synthesis or degradation of RNA is called regulatory RNA. It is interesting to note that Jacob and Monod had initially speculated that the primary product of Repressor of β-galactosidase "may be a polyribonucleotide", although this turned out to be incorrect, as the functional Repressor synthesised is a polypeptide. One of the first discoveries of RNA-mediated gene expression control dates back to 1983. In *E. coli*, Simons and Kleckner discovered that RNA transcribed from the antisense strand of IS10, a Tn10 transposon, inhibits the translation of the transposase encoded in IS10 (Simons and Kleckner, 1983).

The first regulatory RNA discovered in eukaryotes was *Xist* RNA in the mammalian cell. Mammalian female cells possess two X chromosomes, and one of them is randomly inactivated in a process called X chromosome inactivation. The *Xist* gene was identified as a key factor that regulates this phenomenon, but, surprisingly, it has no open reading frame (ORF), suggesting that the *Xist* gene does not code for a protein, and that the RNA transcript itself has regulatory functions (Brockdorff et al., 1992).

In the following year, 1993, Ambros and Ruvkun groups discovered that the *lin-4* gene *C. elegans* encodes a small RNA that binds to the $3'$ untranslated region (UTR) of *lin-14* mRNA

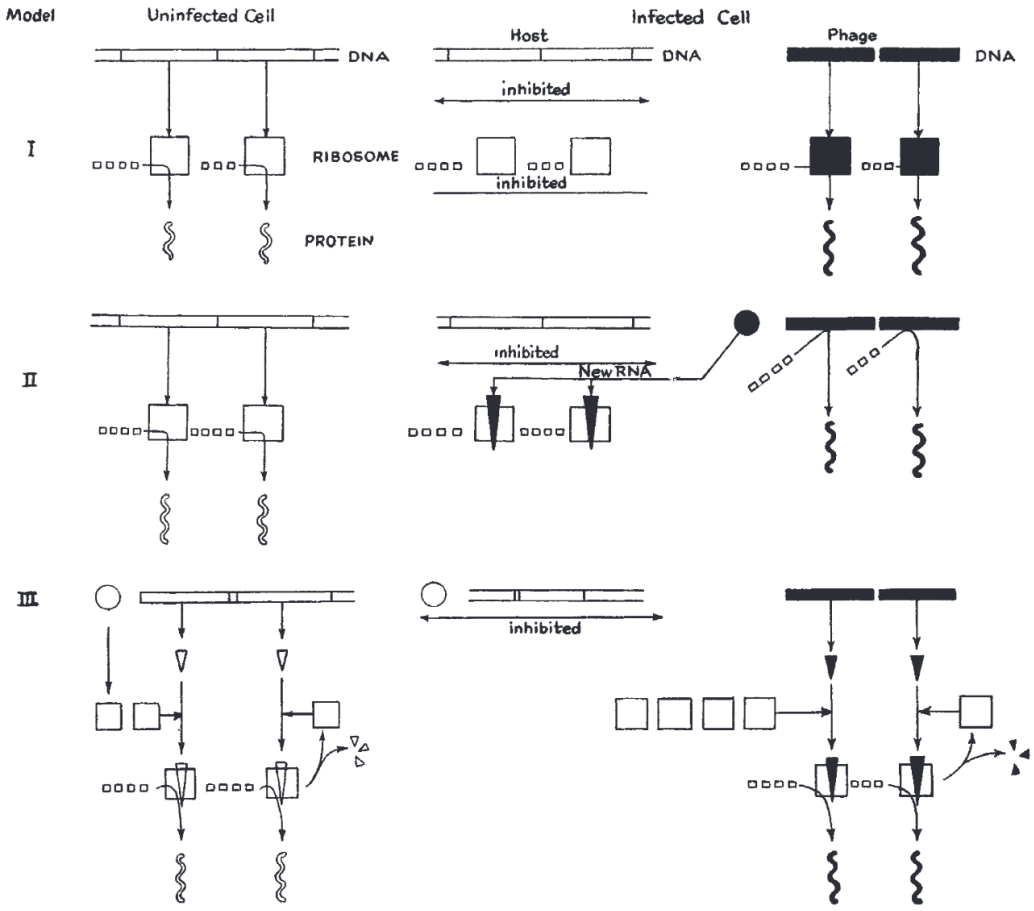


Fig. 1.2 Models of protein synthesis proposed by 1961
Diagrams represent three different protein synthesis models in bacteria in response to phage infection proposed by 1961 (taken from Brenner et al. (1961)). Three models were proposed to explain the fact that after phage infection bacterial cells stop synthesising bacterial proteins, and, instead, phage proteins are synthesised. (I) A specialised ribosome is formed for each gene, and, upon phage infection, phage gene-specific ribosomes are synthesised and bacterial ribosome is somehow inhibited. (II) Some phage-derived RNA binds to the specialised host ribosomes and stalls host protein synthesis. Phage proteins are synthesised on phage DNA. (III) Only non-specialised ribosomes exist in the cell and "messenger RNA" is synthesised from DNA and used as a template for protein synthesis as ribosome. A phage infection inhibits the synthesis of host mRNA, and phage mRNA instead "hijacks" host ribosomes to synthesise phage proteins.

and inhibits the translation of it (Lee et al., 1993; Wightman et al., 1993). This is the first discovery of regulatory small RNA that is later named as microRNA (miRNA).

Since then, numerous research projects have been conducted to explore the biological functions of two major classes of regulatory RNA: long non-coding RNA (lncRNA) and small RNA. Although there is no concrete definition for each term, the former generally refers to RNA longer than 200 nucleotides (nt) and the latter indicates ones that are shorter than this. Some major non-coding RNAs in mammals are summarised in Table 1.1 (Cech and Steitz, 2014; Kim et al., 2009).

Table 1.1 Major mammalian non-coding RNA species

Functional category	Name	Length (nt)
Translation	rRNA (ribosomal RNA)	20-5,000
	tRNA (transfer RNA)	70-90
RNA processing	snRNA (small nuclear RNA)	100-300
	snoRNA (small nucleolar RNA)	70
Gene regulation	siRNA (small interfering RNA)	18-22
	miRNA (microRNA)	21-24
	piRNA (PIWI-interacting RNA)	23-29
	Xist	17,000
	Tsix	40,000
	Hotair	2,000

1.2.1 Small RNA

Small RNA-induced gene regulation is now known as RNA interference (RNAi), and the detailed mechanisms have first been described by Fire et al. (1998). Decades of studies have now revealed complex biogenesis pathways of various small RNAs (Fig. 1.3). In addition to endogenous small RNA such as miRNA, RNAi can also be induced by introducing exogenous double-stranded RNA (dsRNA). This dsRNA can be further processed in the host cells and generate small interfering RNA (siRNA) that induces RNAi on target mRNAs.

Since the development of high-throughput RNA sequencing methods, an increasing number of small RNA species have been discovered. Here, I summarise the details of three major classes of small RNA species: miRNA and piRNA as most well-studied examples, and tRNA-derived fragments as one of the novel small RNAs.

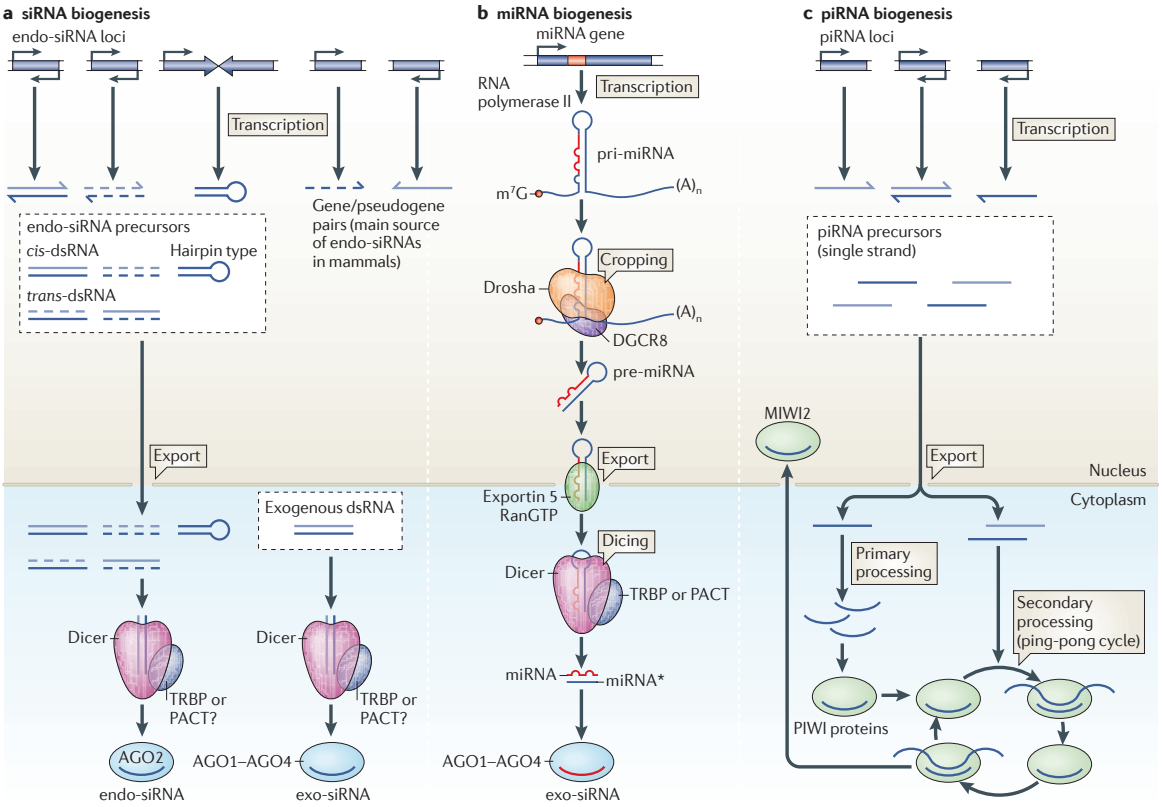


Fig. 1.3 Biogenesis pathways of small RNA species in mice
Schematic representation of the biogenesis pathways of known small RNA species (taken from Siomi et al. (2011)). (a) dsRNA precursors for siRNA can be produced either from endo-siRNA-encoding loci in the genome or exogenously introduced dsRNA. The dsRNA is then diced by Dicer and generates mature siRNA. The siRNA is then bound by an AGO protein to target mRNAs. (b) Pri-miRNA transcribed is bound by Drosha and DGCR8, and processed into pre-miRNA. Pre-miRNA is then exported to the cytoplasm by Exportin-5. Similar to siRNA, pre-miRNA is further cleaved by Dicer to generate mature miRNA and loaded onto an AGO. (c) piRNA precursors are generated as single-stranded RNA and exported to the cytoplasm. The precursors undergo primary processing, followed by secondary processing called "ping-pong cycle" to amplify the piRNAs. The processed piRNAs are imported back to the nucleus and suppress targets with MIWI2. Other organisms such as *C. elegans* possess completely different biogenesis pathways.

microRNA

microRNA (miRNA) is a class of regulatory RNA that is 18-22 nt long and widely conserved from plants to humans. From miRNA genes, long primary miRNA (pri-miRNA) are transcribed by RNA polymerase II (Pol II). Since a miRNA gene normally contains a pair of complementary sequences, pri-miRNA forms a hairpin structure, which is subsequently cleaved off by an enzyme, Drosha. This cleaved-off hairpin, which is 60-70 nt long, is now called precursor miRNA (pre-miRNA). Pre-miRNA is then transferred to the cytoplasm by a protein Exportin and further cleaved by Dicer to generate mature miRNA. One of the strands is loaded on an Argonaute (Ago) complex, which is the effector complex that binds to mRNAs to inhibit their translation or to degrade them.

The targeting mechanism of miRNA has also been investigated extensively. One of the canonical miRNA targeting systems is that a 7-8 nt long sequence at the miRNA's $5'$ end (2nd to 7th nucleotides from its $5'$ end) called "seed" base-pairs with the $3'$ end of mRNA that contains one or more complementary sequences to the seed sequence (Lewis et al., 2005). Although some experiments that identify direct interactions between mRNA and miRNA revealed that there are non-canonical binding independent of seed matchings, follow-up analyses found that miRNA-mediated gene silencing is only induced with canonical pairs. This indicates that seed sequence-mediated targeting is the core targeting mechanism of miRNAs.

Based on the targeting mechanisms of miRNAs through their seed sequence, a number of bioinformatic software have been developed to computationally predict which mRNAs are targeted by a given miRNA (Agarwal et al., 2015; van Dongen et al., 2008). Although multiple aspects, such as the conservation level of the seed region, are taken into consideration in these software, they still yield false-positives and false-negatives when validated experimentally, suggesting that the prediction models still have room for improvements.

piRNA

While miRNAs are expressed ubiquitously, some small RNA species are expressed only in specific cell types. PIWI-interacting RNA (piRNA) is the most well-known such small RNA, and it is believed to be expressed mainly in germ cells. *piwi* (P-element induced wimpy testis) gene was discovered in *Drosophila* through a screen using a transposon called P-element. This assay revealed that *piwi* mutation induces male and female sterility (Lin and Spradling, 1997). The subsequent analyses discovered that *piwi* homologues in various animals are essential for viable germ cells, and that *piwi* proteins are homologues of Argonaute proteins (Batista et al., 2008; Kuramochi-Miyagawa et al., 2004). An immunoprecipitation (IP) assay of mouse

PIWI protein called MIWI revealed that small RNAs that are 26-31 nt in length, which are distinct from previously known small RNAs, are bound by MIWI. Genome-wide sequence analysis, as well as PIWI knock down experiments, revealed that its primary function is to silence transposons to maintain genome integrity (Aravin et al., 2007). However, there are also some piRNAs that regulate host transcripts, suggesting that the functions of piRNA are not limited to transposon silencing.

Different species have completely different piRNA biogenesis pathways. While mammals and *Drosophila* amplify piRNAs using a mechanism called the "ping-pong cycle", in which precursor RNA is cleaved using mature piRNAs as a template (Brennecke et al., 2007; De Fazio et al., 2011), *C. elegans* use RNA-dependent RNA polymerase (RdRP) to achieve piRNA amplification (Das et al., 2008; Lee et al., 2012).

Although piRNAs target mRNAs that contain homologous sequences to them, it is known that the target sequence with a few mismatches can still be targeted by piRNAs (Bagijn et al., 2012). This targeting mechanism makes *in silico* piRNA target prediction challenging and thus requires experimental validations.

tRNA-derived fragment (tRF)

In small RNA-seq data analysis, reads mapping to tRNA or rRNA had been thought to be random fragments from mature tRNA or rRNA and have no gene regulatory roles. However, when a group looked into small RNA-seq reads in human prostate cancer cell lines carefully while maintaining tRNA-mapping reads, they found out that the number of reads mapping to tRNA-coding regions is the second largest after miRNA (Lee et al., 2009). Surprisingly, unlike what normally happens with degraded RNA, the small RNA reads are derived from specific loci of tRNA coding regions, either from $5'$ or $3'$ end, which suggests the presence of an active molecular machinery that breaks tRNA into tRNA-derived small RNA, or tRNA-derived fragment (tRF). Inhibition of tRF by antisense oligos resulted in transcriptional alteration of multiple genes and affected cell viability, suggesting their regulatory potential.

Since then, tRF has been discovered in various cell types, and a growing number of studies have been conducted to explore the link between tRF and gene expression. It started to attract more attention in 2016 when two papers were published in *Science*, both of which reported that tRFs derived from epididymis are transferred to sperm and affect the phenotype of offspring intergenerationally (Chen et al., 2016; Sharma et al., 2016). They reported that dietary alteration affects the repertoire of tRFs in sperm, which is transferred from the epididymal epithelium. This, in turn, alters the transcriptional dynamics in embryos, leading to phenotypic differences in offspring. However, the mobility of tRF was only suggested by showing the correlation between tRFs expressed in different parts of epididymis and in

sperm collected from the corresponding parts of the epididymis. Deeper investigations are needed to clearly prove their intercellular mobility. tRFs as well as these tRF-related reports are further discussed in Chapter 4.

1.2.2 Long non-coding RNA

High-throughput RNA analyses also identified numerous lncRNA genes across the genome in different organisms. lncRNAs can arise from intergenic regions, or intronic or exonic regions of another gene. Some are conserved across organisms and others are species-specific. Some are polyadenylated and others are not. Unlike small RNA species, there is no known common features among lncRNAs. Although the first regulatory RNA, *Xist* RNA, showed a very clear molecular phenotype involved in X inactivation, studies on functions of lncRNAs have not always yielded clear insights into their biological roles. Recent large-scale knock-out (KO) experiments on zebrafish and *C. elegans* showed that many lncRNA KO animals are viable and show no clear phenotypes (Akay et al., 2019; Goudarzi et al., 2019). Below, some of the well documented regulatory lncRNAs are summarised.

Xist RNA

One of the most studied and characterised lncRNA is Xist, which is encoded on the X chromosome of eutherian mammals. In female cells, where two X chromosomes are present, it is known that one of the X chromosomes is silenced for dosage compensation. Studies have identified that *Xist* RNA synthesised from the *Xist* gene on the X chromosome spreads across one of the X chromosomes through 3D chromatin architecture (Engreitz et al., 2013) and silences it by recruiting repressive chromatin modifiers such as the Polycomb complex through a repeat region called RepA (Chu et al., 2015; Zhao et al., 2008).

HOTAIR RNA

The *Hox* clusters are important for development, and a non-coding gene named *Hotair* (Hox transcript antisense RNA) was discovered in the *HoxC* cluster in mice. A paper reported that a deletion of *Hotair* induced transcriptional changes in trans and mouse *Hotair* KO mice showed skeletal malformations (Rinn et al., 2007). *Hotair* RNA was later shown to be associated with the Polycomb repressive complex 2 (PRC2) and to recruit it to the genomic target regions (Li et al., 2013).

These results, however, have been controversial. Reanalysis of the same mouse *Hotair* KO showed little phenotypic changes in another group (Amândio et al., 2016). Also, the Hotair-binding regions in the genome identified by Li et al. do not significantly overlap

with the previously reported PRC2-binding regions. This controversy is further discussed elsewhere (Selleri et al., 2016), but no clear conclusion has been reached yet.

1.3 RNA biogenesis

1.3.1 Biosynthesis

RNA consists of a ribose sugar chain and bases. There are four bases in RNA: adenine (A), uracil (U), cytosine (C), and guanine (G). The purines (A and G) and the pyrimidines (U and C) have different biosynthetic pathways, but both require the same molecule, phosphoribosyl pyrophosphate (PRPP), which is a derivative from ribose-5-phosphate produced from the pentose phosphate pathway.

For purine synthesis, PRPP is converted to inosine monophosphate (IMP). Adenine monophosphate is synthesised directly from IMP, while guanine monophosphate (GMP) is synthesised through an intermediate, xanthosine monophosphate (XMP) (Fig. 1.4).

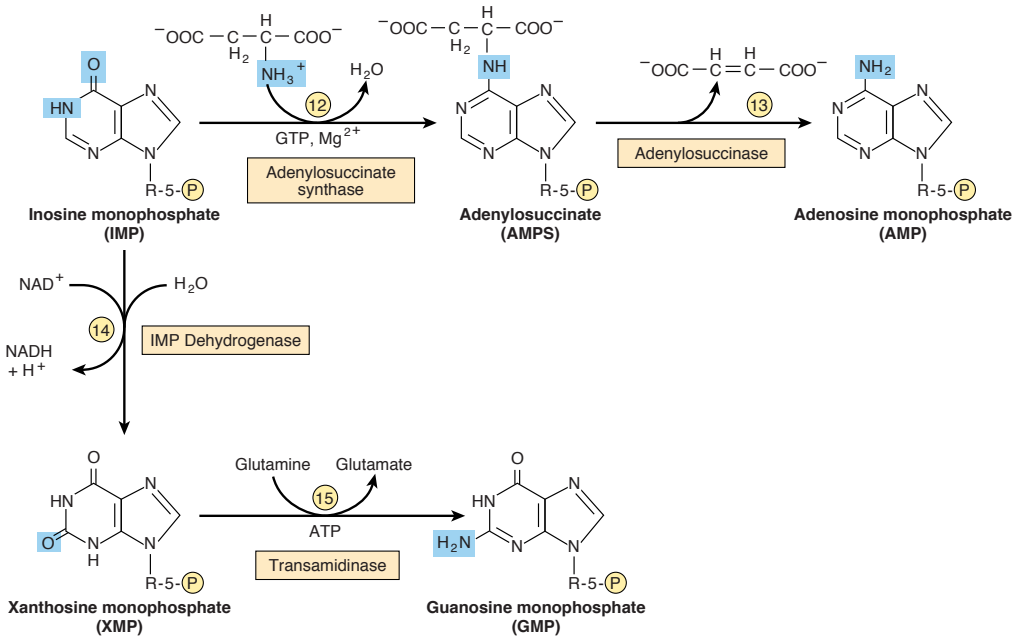


Fig. 1.4 Purine NMP biosynthetic pathway
Biosynthetic pathways of GMP and AMP from IMP (taken from Murray et al. (2009, p. 295)).

On the other hand, pyrimidine nucleotides synthesis begins with the synthesis of uridine monophosphate (UMP) from orotate and PRPP (Fig. 1.5). UMP is then phosphorylated to

generate uridine diphosphate (UDP) and uridine triphosphate (UTP). UTP is then converted to cytosine triphosphate (CTP).

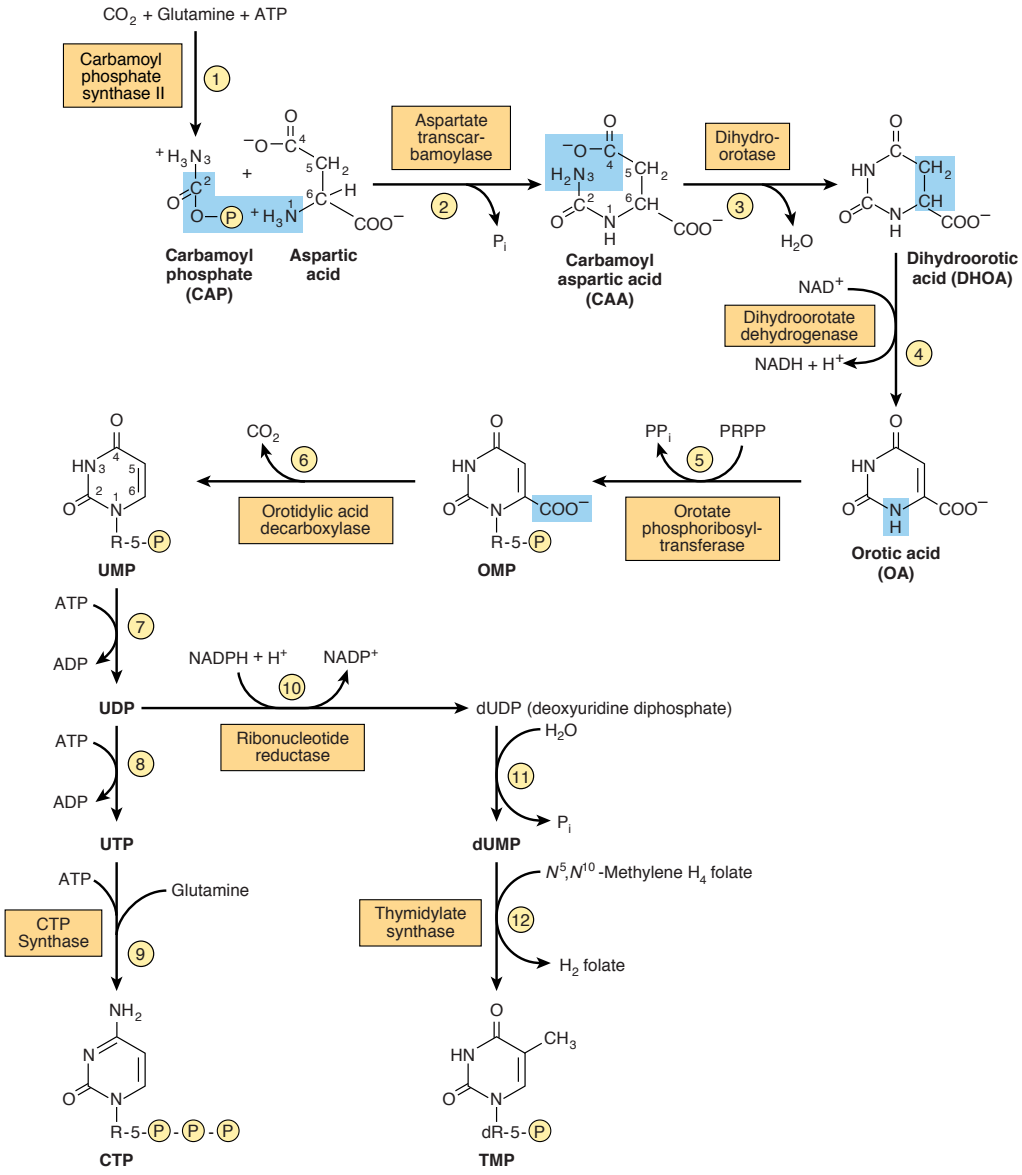


Fig. 1.5 Biosynthetic pathway of pyrimidine nucleotides
Taken from Murray et al. (2009, p. 298).

In addition to these *de novo* synthesis of nucleoside triphosphates (NTPs), cells also are able to salvage intermediate molecules from degraded RNAs. Purine bases are salvaged by phosphoribosyltransferases, which add PRPP to nucleobases to create nucleoside monophosphates (NMPs). Adenine phosphoribosyltransferase (APRT) and hypoxanthine-guanine phosphoribosyltransferase (HGPRT) are involved in adenine and guanine salvage,

respectively. Although uracil phosphoribosyltransferase (UPRT) in bacteria is known to salvage uracil to UMP (Pfefferkorn, 1978), mammalian UPRT does not bind to uracil, and no enzymatic activity was detected (Cleary et al., 2005; Li et al., 2007). Instead, uridine and cytidine are salvaged to UMP and cytidine monophosphate (CMP), respectively, by uridine-cytidine kinases.

1.3.2 Transcription

The eukaryotic mRNA transcription consists of multiple steps (Fig. 1.6). When expression of a gene is required, a pioneering transcription factor (TF) binds to a region proximal to the core promoters to open up the chromatin. General transcription factors (GTFs) and RNA polymerase II (Pol II) are recruited to the core promoters, which include several DNA sequence motifs, to initiate transcription. Pol II joins NTPs in an order defined by the template DNA sequence through the complementary base-pairing: A is paired to T, G to C, and C to G. The only exception is U, which is paired to A instead of T. The transcribed RNA undergoes the maturation processes, such as a cap addition and polyadenylation, and is transferred to the cytoplasm for translation.

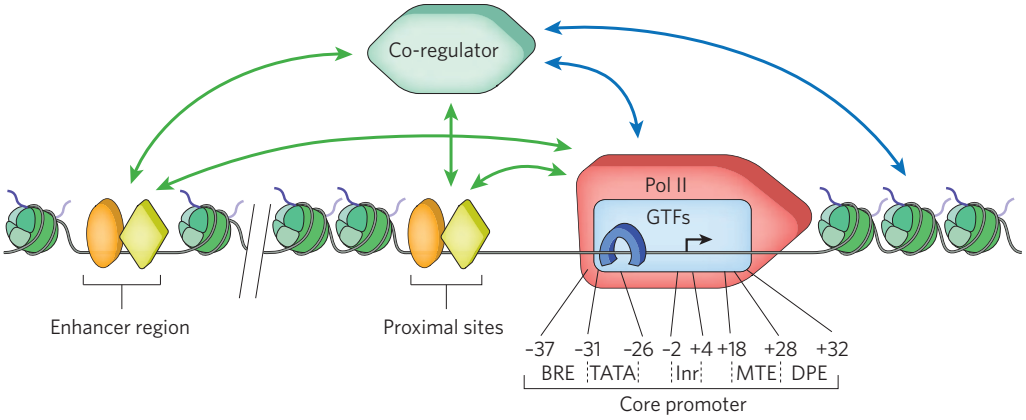


Fig. 1.6 Sturcture and interactions of eukaryotic promoter
Pioneering TFs (orange and yellow) bind to either the loci near the promoter or the enhancer region to make the chromatin accessible. Pol II and GTFs are recruited to the accessible core promoter. Several consensus motifs are known to be present in the core promoter, such as the B recognition element (BRE), the TATA box (TATA), the initiator (Inr), the motif ten element (MTE), and the downstream promoter element (DPE). Not all of these motifs may be present within a promoter. The locations of the motifs relative to the transcription start site (black arrow) are also shown. Known interactions are shown in green and blue arrows. Figure taken from Fuda et al. (2009).

1.4 Metabolic RNA labelling methods

To study the dynamics of RNA transcription, various nucleotide analogues or NTP analogues have been employed to date. These analogues contribute to the cellular NTP pool and then used to synthesise RNA, making the newly-synthesised RNA distinct from pre-existing RNAs. Since U only exists in RNA but not in DNA, uridine or UTP analogues have often been used to study RNA dynamics. Here, I review different analogues used so far and their advantages and limitations.

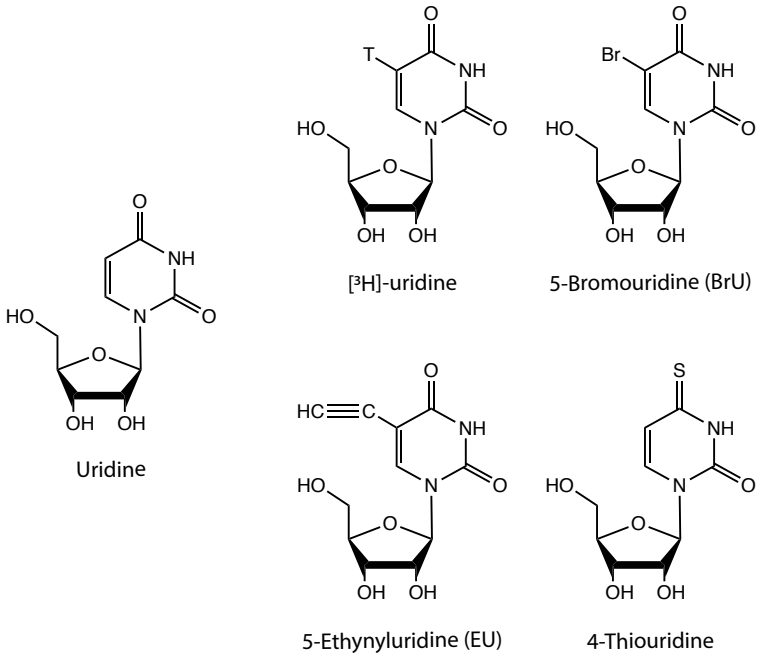


Fig. 1.7 Chemical structures of uridine and commonly-used uridine analogues

1.4.1 Radioautography

At the beginning of RNA labelling research, radioisotope-containing nucleotides, such as [^{3}H]-uridine (Uddin et al., 1984), were used. After exposing animals or cells to an analogue for a limited time, the tissues or cells are examined by radioautography. One can visually assay the abundance of newly synthesised RNA by quantifying the number of radioactive granules detected, and which cells or subcellular compartments contain newly transcribed RNAs. The major advantage of this method is that the radiolabelled nucleotide should have little effects on the functions of RNA molecules, as there is no structural differences

from the native nucleotides. However, most research institutes nowadays have very strict regulations on radioactive agents, and thus conducting an experiment using radioactive nucleotides is not straightforward. Also, since there is currently no established method to isolate the radiolabelled RNA, the sequence of labelled RNA cannot be examined unless a new sequencer that utilises single-molecule detectors will be improved to detect radiolabelled nucleotides.

1.4.2 Antibody detection

5-Bromouridine (BrU) has become a more common labelling agent because of its easier and more convenient usage. BrU can be detected by a specific antibody against it (Wansink et al., 1993). Thus, RNA containing BrU can be visualised by immunostaining, and its cellular/subcellular location can be studied. The abundance of newly transcribed RNA can also be quantified based on the fluorescence intensity. The availability of anti-BrU antibody enables specific pull-down of RNA containing BrU, and by combining with it either a low- or high-throughput RNA quantification method, one can quantify the abundance of each labelled RNA transcripts (Kageyama et al., 2004).

1.4.3 Click chemistry

As an alternative to the antibody-based detection of labelled nucleoside, 5-ethynyluridine (EU) has been employed. EU can be visualised with fluorescent azides through a Sharpless–Meldal copper (I)-catalysed Huisgen cycloaddition reaction, which is also referred to as "click" chemistry (Jao and Salic, 2008; Rostovtsev et al., 2002; Tornøe et al., 2002). Since this reaction does not require cell permeabilisation, EU can also be used for whole-mount staining (Jao and Salic, 2008), which cannot be achieved with BrU. Also, through click chemistry with biotin, EU-labelled RNA can be pulled down with streptavidin beads, which allows direct detection and quantification of the labelled transcripts just like the antibody pull-down with BrU. A major limitation with EU is that the attachment of molecules with click chemistry is irreversible. Hence, the isolated RNA is not suitable for cDNA synthesis, and EU is not a suitable agent to be combined with next generation sequencing approaches.

1.4.4 Reversible disulfide chemistry

Recently, 4-thiouridine has been widely used to study the dynamics of RNA. Similar to EU, 4-thiouridine-containing RNA can be pulled-down with thiol-group specific biotinylation, followed by streptavidin isolation of the biotinylated RNA (Cleary et al., 2005; Dölken et al.,

2008). The major benefit of 4-thiouridine over EU is that the attached biotin can be removed in the elution step by breaking the disulfide bond with a reducing agent, and thus the eluted RNA can be used for high-throughput RNA sequencing (Kenzelmann et al., 2007). Different methods that employ 4-thiouridine for metabolic labelling are further discussed in Chapter 3.

1.5 High-throughput RNA sequencing (RNA-seq)

In studying RNA biology, it is essential to quantify the amount of each RNA species in a given sample accurately. Owing to the development of high-throughput DNA sequencing (HTS) methods, accurate and simultaneous quantification of multiple RNA species has become possible. Here, one of the most employed strategy for high-throughput sequencing, the Illumina method, is summarised, and the benefits and limitations in RNA quantification are discussed.

1.5.1 Illumina DNA sequencing

With the traditional DNA sequencing methods, the major limitation to its throughput is that the reaction is carried out in each well of 96-well plates. To overcome this limitation, massively-parallel DNA sequencing methods have been invented. Here, one of the most frequently used methods, the Illumina method, is summarised (Bentley et al., 2008) (Fig. 1.8). DNA fragments are first attached with short oligonucleotides called adapters and captured on a flat reaction chamber that have a number of oligonucleotides that are complementary to the adapter sequences on its surface. Then, libraries are amplified to form clusters that are originated from the individual DNA molecule through bridge-PCR. The DNA templates are sequenced by base-by-base incorporation of reversible terminators, 3′-*O*-azidomethyl 2′-deoxynucleoside triphosphates with the four bases (A, T, G, and C), which are labelled with a different fluorophore. The sequencing can be performed from only one end (single-end sequencing) or from both ends (paired-end sequencing).

1.5.2 Sequencing error

Although HTS is a powerful method in determining a sequence of nucleotides, it still has inherent sequencing errors, as with other sequencing methods. Several potential sources of errors have been suggested: misincorporation of nucleotides during PCR, preferential incorporation of certain ddNTP (Schirmer et al., 2016), and insufficient removal of terminators (Pfeiffer et al., 2018). As a result, Illumina Hiseq systems are shown to have an order of 0.1% of the misreading of bases.

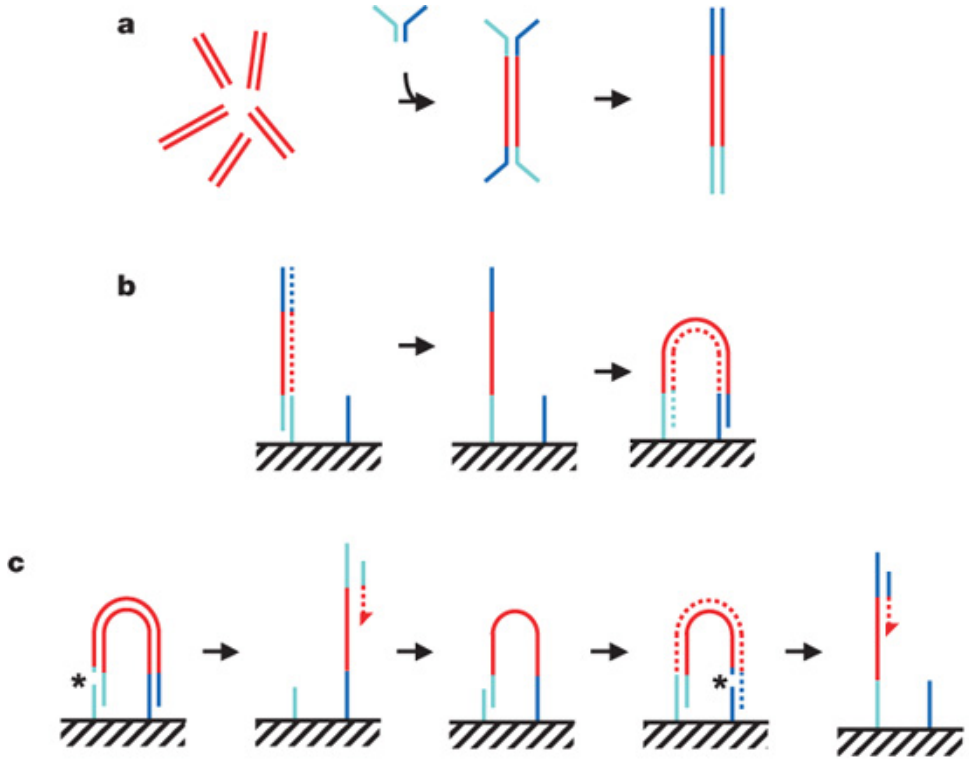


Fig. 1.8 Schematic of Illumina DNA sequencing
(a) Adapters (shown in blue and light blue) are ligated to fragmented DNA. (b) Denatured DNA molecules are annealed to oligonucleotides on the flowcell surface that are complementary to the adapters. A new sequence (shown as dotted line) is synthesised from the original sequence by extension. The synthesised strand is then annealed to another oligonucleotide that is complementary to the other adapter on the other side by forming a bridge, and a new strand is synthesised (dotted line). This series of amplification (bridge PCR) is repeated until each cluster reaches 10 µm in diameter. (c) For sequencing, the amplified libraries are cleaved at one adapter (asterisk) and linealised. Sequencing run is conducted by a sequencing primer complementary to the adapter. For paired-end sequencing, the second strand is resynthesised by bridge PCR, and the sequence is determined by a primer complementary to the other adapter. Figure taken from Bentley et al. (2008).

There are also methods to alleviate this error rate. Base-calling software assigns each base a quality score called phred score to estimate the accuracy of base calling (Ewing and Green, 1998; Ewing et al., 1998). Since the bases with a low phred score are more likely called wrong, these bases can be removed for improved base-calling. Also, it is suggested to add a short stretch of nucleotides to either $5'$ or $3'$ end of the target DNA (Fu et al., 2011; Hug and Schuler, 2003). Since all the fragments originating from the same DNA molecule will have the same "tag", PCR error can be mitigated by finding a consensus sequence among the reads with the same tag. Moreover, paired-end sequencing can also be used to increase the accuracy of base-calling. Since the same molecule is sequenced twice, the probability of calling a base wrong would decrease significantly.

1.5.3 RNA-seq

Since RNA can be reverse transcribed into cDNA, HTS can also be used to determine the sequence of RNA molecules (Wilhelm and Koopman, 2006). Furthermore, based on the number of detected sequences for a given RNA species, relative quantification of each transcript can be achieved.

One of the most frequent applications of RNA-seq is discovery of transcriptomic change between two samples. Since the read count for each gene is only a relative measure of transcript abundance in each sample, the normalisation method is the key to accurately identify differentially expressed genes. A number of bioinformatic methods to normalise read count have been invented and widely applied (Anders and Huber, 2010; Bray et al., 2016; Trapnell et al., 2010). Also, instead of solely relying on an *in silico* approach, addition of a small amount of RNA with distinct sequence from sample RNA prior to RNA-seq library prep was also proposed (spike-in) (Zook et al., 2012). Since the ratio of spike-in RNA to sample RNA is known for each sample, the number of reads for each gene can be adjusted to the number of reads obtained from spike-in RNA. Another normalisation strategy is to add a short stretch of DNA (UMI: unique molecular identifier) to cDNA before PCR amplification. With this modification, the number of RNA molecules can be quantified by counting the number of unique UMIs for each gene.

1.5.4 Single-cell RNA-seq (scRNA-seq)

Normally, RNA-seq was often performed using RNA extracted from a number of cultured cells or a part of a tissue from an animal. Thus, the gene expression quantified from this method is an average among the cells from which the RNA is extracted. To obtain the transcriptional landscape of distinct individual cells, a more sensitive RNA-seq method that

enables to prepare libraries from RNA extracted from single cells has been developed (Tang et al., 2009). This scRNA-seq not only enables to obtain transcriptomic variance among cells, but also to identify distinct cell populations within given animal tissues, or to differentiation trajectories within an animal (La Manno et al., 2018). To interpret and cluster high dimensional data of gene expression for each cell, it is often used to perform dimensionality reduction methods. Principal component analysis (PCA) (Pearson, 1901) or *t*-distributed neighbour embedding (*t*-SNE) (Maaten and Hinton, 2008) can project high-dimensional data to lower dimensional space (e.g. 2D or 3D) to visualise cells with similar gene expression patterns. By using known maker genes, one can orientate which cell type is represented by a given cluster and can find unknown cell populations. However, great care must be taken since the new clusters can sometimes arise from technical artefacts (O'Flanagan et al., 2019; van den Brink et al., 2017). Studies have shown that cell dissociation methods can induce differential gene expression in a part of cells, which was miscalled as a new cluster.

1.6 SLAMseq

Although RNA-seq is a powerful tool to quantify the number of RNA molecules existing in the cell, it still lacks temporal information, and thus metabolic labelling has still been the best way to study transcriptional dynamics. To overcome the drawbacks of biotin-streptavidin isolation system, an alternative strategy for thio-labelled RNA discovery called thiol(SH)-linked alkylation for the metabolic sequencing of RNA (SLAMseq) was developed (Herzog et al., 2017). Instead of isolating the labelled RNA from a total RNA pool, an alkylating agent, iodoacetamide (IAA), is used to attach a chemical residue to the thiol group specifically (Fig. 1.9A). This alkylated 4-thiouracil induces mispairing of bases in the reverse transcription (RT) step in RNA-seq library preparation: G is base-paired to this aklylated uracil instead of A (Fig. 1.9B). Thus, after polymerase chain reaction (PCR) and high-throughput sequencing, reads generated from the labelled RNA contains thymine-to-cytosine (T>C) conversions at the positions where 4-thio-UTP was incorporated. T>C-aware read aligner, which was specifically developed for SLAMseq analysis, aligns reads with T>C mismatches to the genome for a sensitive discovery of T>C conversions.

1.6.1 SLAMseq data analysis

Since the reads obtained from SLAMseq contain more frequent T>C mismatches compared with conventional RNA-seq methods, read aligner developed for standard RNA-seq cannot map reads to the genome due to their high mismatch rate. Thus, SLAMseq requires an

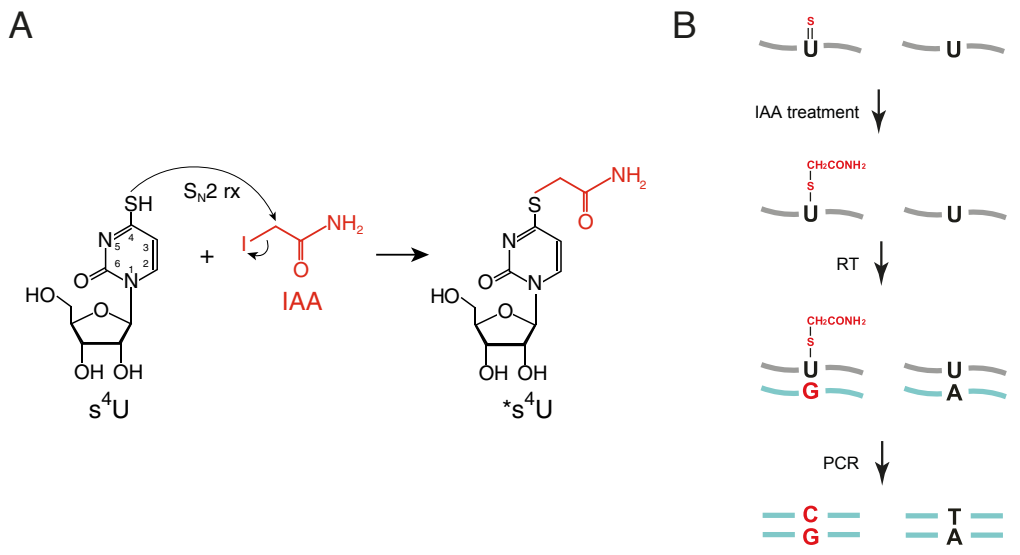


Fig. 1.9 SLAMseq biochemistry
(A) Alkylation reaction between IAA and the thiol group of 4-thiouridine is shown (taken from Herzog et al. (2017)). A carboxyamidomethyl group is attached to the thiol group of 4-thiouridine (s^4U) through a nucleophilic substitution (S_N2) reaction. (B) SLAMseq workflow for thiol-containing and non-containing RNA is shown. The chemical residue added thorough the alkylation reaction induces mispairing of G instead of A in the RT step and leads to T>C base conversions in reads resulting from it.

optimised strategy to align reads to the genome for an accurate estimation of gene expression and T>C ratio. To achieve a better handling of reads containing frequent T>C mismatches, Neumann et al. have developed a new software called SLAM-DUNK, which consists of four consecutive steps: "map", "filter", "snp", and "count" (Fig. 1.10).

In the "map" step, reads are mapped to the genome allowing T>C mismatches. This is achieved by modifying an existing aligner, NextGenMap (Sedlazeck et al., 2013), which was originally developed for mapping reads to a genome that is highly polymorphic. In determining where to align a read to the given genome, the native algorithm of NextGenMap assigns a penalty score when there is a mismatch or a gap between the read and the mapped genomic locus. SLAM-DUNK's "map" step was configured not to penalise a mismatch when it is a T>C mismatch in order to maintain reads mapped to the genome with high T>C mismatches.

In the "filter" step, reads with a low mapping score and those that are ambiguously mapped are removed. If a read cannot be assigned to a unique location, the software uses supplied transcript annotation information to resolve this ambiguity. If a read is mapped to multiple genomic loci, and only one of them overlaps with the annotation, only the location in the annotation is assigned to the read. If a read is mapped to multiple genomic regions that are included in the annotation, or does not overlap with any of the annotated loci, the read is removed from the analysis.

Since the T>C mismatches found in the reads can also be derived from single-nucleotide polymorphisms (SNPs), the "snp" step identifies T>Cs that are likely derived from SNPs. In diploid cells, a SNP in the genome can in theory result in substantial transcripts containing the SNP unless imprinted, whereas the intracellular concentration of 4-thiouridine can only allow transcripts with <10% of 4-thiouridine incorporated for each T position. Thus, T positions with a higher T>C conversion rate than the specified threshold (25% for diploid cells) are considered to be derived from SNPs rather than 4-thiouridine incorporations and are removed from the downstream T>C counting.

Finally, in the "count" step, the T>C mismatches that were not filtered out in the previous "filter" step are counted for each gene in the annotation file. A table summarising T>C conversion rate for each gene is also generated as an output. Since sequencing errors can also generate low-rate T>C conversions that are not removed in the previous step, T>C containing bases with a Phred quality score higher than the threshold are included for this count.

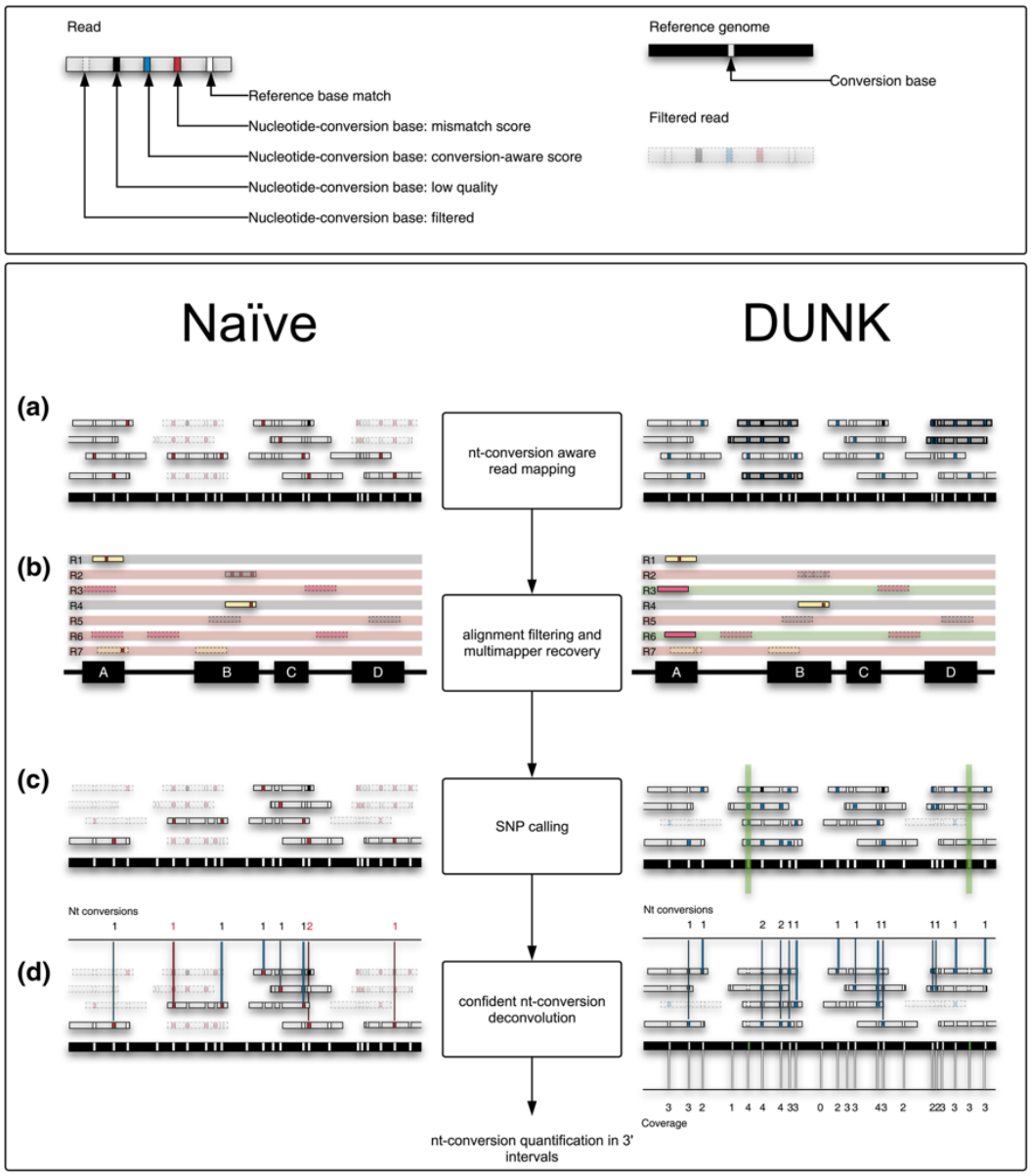


Fig. 1.10 Comparison of read mapping strategies with and without SLAM-DUNK software

Four analysis steps in SLAM-DUNK are shown to illustrate how the software can better handle reads with T>C mismatches compared with when only a standard read aligner is used (naive) (taken from Neumann et al. (2019)). (a) In the "map" step, SLAM-DUNK maintains reads with T>C mismatches that are filtered out in a naive aligner (shaded reads in the left pane). (b) In the "filter" step, when a read is mapped to a unique locus in the genomic region that is included in the annotation file provided, the read is mapped there even though it is mapped to other genomic loci that are not in the annotation. Such reads could be discarded with a naive mapper (R2, R6). (c) In the "snp" step, if the frequency of T>C mismatch in a given locus is higher than a threshold, this locus is removed from T>C counting. This step is efficient in removing T>C mismatches that are derived from SNPs. (d) In the counting step, T>C mismatches that were not filtered in the previous step are counted. When compared with a naive strategy, SLAM-DUNK detects T>Cs more sensitively and accurately.

1.7 Outline and aims of this thesis

RNA has been known as a key biological molecule, and it exists in all known living cells. Although, since its discovery in the 1960s, numerous research projects have been conducted to assess its significance in cellular processes, and molecular machineries producing RNA have been discovered, the complete picture of RNA dynamics is still elusive.

Metabolic RNA labelling has been a powerful tool to study RNA dynamics. Although different nucleotide analogues have been used so far, 4-thiouracil and 4-thiouridine have become one of the most commonly used analogues due to their convenience, minimal structural difference from the native nucleotides, and availability of efficient detection methods. During my PhD, I have developed a novel *in vivo* metabolic RNA labelling method through the collaboration with a group in Austria and used the method to answer various biological questions. Below, I summarise the aims I achieved during my PhD.

- To establish a novel metabolic RNA labelling method in a specific type of cells *in vivo* (SLAM-ITseq).
- To test if there are any endogenous RNAs that are mobile between different types of cells in *M. musculus* using SLAM-ITseq.
- To identify transcripts synthesised from zygotic genome using metabolic labelling.

In Chapter 3, the development of SLAM-ITseq is shown. By labelling RNAs in well-studied cell types, sensitivity and specificity of the method in multiple cell types are analysed. Using this novel *in vivo* RNA labelling method, intercellular mobility of RNA is assayed in Chapter 4. By generating multiple mouse strains with RNA labelled in different cell types, RNA mobility between different cell types as well as RNA released into circulation are comprehensively assayed. In Chapter 5, using the same detection method, transcriptional dynamics in early mouse preimplantation embryos are studied. Stage-specific exposure of embryos to 4-thiouridine enables to identify genes and transposable elements actively synthesised in the embryo.

In summary, through the development and application of novel RNA labelling methods, this thesis explores RNA dynamics in *M. musculus*: cell-type-specific transcriptome *in vivo*, intercellular RNA transfer, and zygotic genome activation in the early preimplantation embryo. Since these questions are all challenging to address with conventional RNA analysis methods, this thesis not only proposes a new tool to tackle these questions, but also provides first direct results addressing each problem.

Chapter 2

Materials and methods

2.1 *M. musculus* methods

2.1.1 Mouse husbandry

All mice were maintained in a specific-pathogen-free facility with sentinel monitoring at standard temperature (19-23°C) and humidity (55 $\pm$ 10%), on a 12 h dark, 12 h light cycle (lights on 7:30–19:00) and fed a standard rodent chow (9% crude fat content, 21% kcal as fat, 0.276 ppm cholesterol, LabDiet). Both food and water were available *ad libitum*. The mice were housed in groups of 3-4 mice per cage in individually ventilated caging receiving 60 air changes per hour. In addition to bedding substrate (Aspen), standard environmental enrichment of a nestlet and a cardboard tunnel were provided. All animals were regularly monitored for health and welfare concerns, and were additionally checked prior to and after procedures. The care and use of mice in the study was carried out in accordance with UK Home Office regulations, UK Animals (Scientific Procedures) Act of 1986 under a UK Home Office license that approved this work (PF8733E07), which was reviewed regularly by the Wellcome Sanger Institute Animal Welfare and Ethical Review Body.

2.1.2 Genetic crosses

All imported mice used were crossed with C57BL/6NTac at least once before used for further crosses. The detailed information about the transgenic mice used in this study is shown in Table 2.1. Homozygous *UPRT* mice (*uprt*/*uprt*) were crossed with hemizygous *Cre* mice (*cre*/*cre*). In the F1 generation, an approximate 50:50 ratio of the Cre^{+} mice (*uprt/0*; *cre/0*) and Cre^{-} mice (*uprt/0*; +/+) were obtained. When mice reached the 10-day age, ear clips were obtained by an ear puncher by Sanger Institute Research Support Facility (RSF) staffs

for animal identification and genotyping. Collected ear clips were stored at -20°C until DNA was extracted as described in a section below.

2.1.3 Tamoxifen administration

Spink8$^{tm1(EGFP/Cre/ERT2)Wtsi}$ transgene generates Cre-ERT2, which is a fusion protein of Cre recombinase and a mutated ligand binding domain of the human estrogen receptor (ER) (Feil et al., 1997). Cre-ERT2 requires 4-hydroxy-tamoxifen (OHT) or tamoxifen to induce Cre recombinase activity (Indra et al., 1999). 20 mg/ml tamoxifen (Sigma-Aldrich) solution in corn oil was made by dissolving tamoxifen in corn oil (Santa Cruz Biotechnology), and the solution was placed in a 37°C incubator overnight with continuous mixing. Both *Spink8-Cre*$^{+}$ and *Spink8-Cre*$^{-}$ mice received 20 mg/ml tamoxifen at 3.75 µl/g body weight dose by i.p. injections for 5 consecutive days with 24 h intervals. These mice were used for 4-thiouracil injections 2 weeks after the last tamoxifen injection.

2.1.4 Administration of RNA labelling agents

4-thiouridine (Sigma-Aldrich) was dissolved in dimethyl sulfoxide (DMSO; Sigma-Aldrich) at 400 mg/ml concentration and further diluted in corn oil (Santa Cruz Biotechnology) in a 1:4 ratio (100 mg/ml final 4-thiouridine concentration). Either the 4-thiouridine solution or DMSO control solution was then i.p. injected into wild-type C57BL/6NTac mice at a dose of 8 µl/g body weight three times every 24 h.

4-thiouracil (Sigma-Aldrich) was dissolved in DMSO at 200 mg/ml concentration and further diluted in corn oil in a 1:4 ratio (50 mg/ml final 4-thiouridine concentration). The 4-thiouracil solution was then injected to both Cre^{+} and Cre^{-} mice at a dose of 8 µl/g body weight three times every 24 h. The mice were culled, and blood and tissues of interest were collected as described below.

2.1.5 Serum collection by cardiac puncture

The mice were anaesthetised with isoflurane 6 h after the last injection and blood was collected by cardiac puncture using a 25G x 5/8" needle (Terumo) and a 1-ml syringe. Cervical dislocation was performed to confirm their death. The collected blood was incubated at room temperature for 30 min and subsequently at 4°C for 30 min. The coagulated blood was then centrifuged at 2,000g for 10 min at 4°C. The supernatant was dispensed into a new 1.5-ml tube, centrifuged at 2,000g for 10 min at 4°C, and the resulting supernatant was

Table 2.1 Mouse strains used in this thesis

Strain	Genotype	Strain ID	Description
Tie2-Cre	Tg(Tek-cre)1Ywa/J	JAX #008863	Endothelial specific Cre line
Vil-Cre	Tg(Vil1-cre)997Gum/J	JAX #004586	Intestinal epithelium specific Cre line
Adipoq-Cre	Tg(Adipoq-cre)1Evdr/J	JAX #010803	Aipocytes-specific Cre line
Spink8-Cre	$\text{Spink8}^{\text{tm1(EGFP/Cre/ERT2)Wtsi}}$	MGI:5633845	Epididymis epithelium-specific Cre line
UPRT	Tg(CAG-GFP,-Uprt)985Cdoe/J	JAX #008863	Cre-inducible UPRT line

recovered again to completely remove blood cells. The obtained serum was stored at -80°C until used for downstream RNA extraction.

2.1.6 Solid tissue collection

Solid tissues (brain, eWAT, intestine, and epididymis) were cut into pieces less than 5 mm in thickness and submerged in 1-ml RNAlater (Sigma-Aldrich) for storage at -20°C. Details of the collection method of each tissue is summarised below.

Brain

The mouse skull was cracked open with surgical scissors, and the whole brain including cerebellum was collected. In order not to exceed the upper limit of RNA concentration that the acid guanidinium thiocyanate reagent, TRIsure (Bioline), can dissolve, only the left hemisphere was used for RNA extraction. The weight of the tissues that TRIsure reagent can dissolve is shown in the manufacturer's instructions. Further details about RNA extraction are described in the next section.

Intestine

Duodenum was collected by incising at the junction of the intestine and stomach, and the 20 mm distal part of it. The luminal side of the duodenum was washed with phosphate-buffered saline (PBS) before submerged in RNAlater.

Epididymal white adipose tissue (eWAT)

Only the right eWAT was collected. The junction with the testis was identified and eWAT was excised at this point.

Epididymis

The caudal and caput parts of the epididymis were collected (Fig. 2.1). Since cauda was also used to isolate sperm, it was stored for RNA extraction after being incubated in M2 medium for sperm collection (details described below).

2.1.7 Sperm and epididymosome isolation

Sperm and epididymosomes were isolated from cauda epididymides as previously described with minor modifications (Sharma et al., 2016). 4 ml M2 medium (Sigma-Aldrich) was

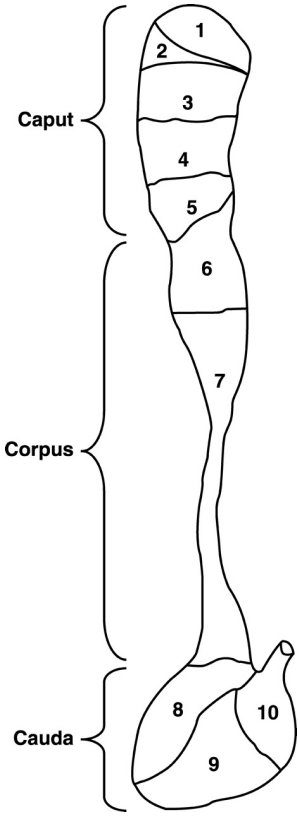


Fig. 2.1 Schematic representation of mouse epidiymal segments
The caput and cauda epididymis were collected based on this scheme (taken from Johnston et al. (2005)).

aliquoted in 5-ml round-bottom polystyrene tubes (Corning) and pre-warmed in a 37°C water bath. Cauda epididymides collected from mice were submerged in the pre-warmed medium after a few incisions were made and incubated at 37°C for 30 min. Epididymides were preserved in 1 ml RNAlater for RNA isolation, and 3 ml supernatant was collected in 1.5-ml tubes and centrifuged at 2,000g for 2 min. The resulting supernatant was dispensed into new 1.5-ml tubes for epididymosome isolation, and the pellet was used for sperm isolation.

Sperm purification

The sperm pellet was washed with 500 µl PBS. The sperm suspension was centrifuged at 2,000g for 2 min, and the supernatant was discarded. The pellet was then resuspended with somatic cell lysis buffer (0.1% SDS and 0.5% Triton-X (Sigma-Aldrich) in nuclease-free water), and incubated on ice for 10 min. The solution was centrifuged at 2,000g for 2 min, and supernatant was discarded. The pellet was washed with 500 µl PBS again followed by centrifugation, and the resulting sperm pellet was stored at -80°C until used for downstream RNA isolation.

Epididymosome isolation

Epididymosomes were isolated using a method that was previously described elsewhere (Sharma et al., 2016). Epididymosome-containing supernatant was centrifuged at 120,000g at 4°C for 2 h, washed with ice cold PBS, and centrifuged again at 120,000g at 4°C for 2 h. The supernatant was discarded, and the epididymosome pellet was used for downstream RNA isolation.

2.1.8 Superovulation and embryo collection

Superovulation and embryo collection were performed following the protocol described elsewhere (Doe et al., 2018) with the support of Evelyn Grau at RSF. 4-6 weeks old female mice were i.p. injected with pregnant mare serum gonadotropin (PMSG; 5 IU, Intervet) and, 48 h later, with human chorionic gonadotropin (hCG; 5 IU, Intervet). After the hCG injection, the female mice were housed with stud male mice overnight for mating. After confirming a copulation plug, female mice were euthanised, and cumulus complex mass was collected from their ampulla. One vial of hyaluronidase (Sigma-Aldrich) was added to the cumulus complex on a dish, and the dish was swirled to disperse the cumulus cells. The embryos were washed with flushing holding medium (AMS Biotechnology), and the fertilised zygotes with two pronuclei were placed into potassium-supplemented simplex optimised medium (KSOM; AMS Biotechnology) that was pre-warmed at 37°C.

2.1.9 Embryo culture

In each dish, 40-60 embryos were cultured in 1 ml KSOM medium at 37°C. 100 µl of 50 mM 4-thiourdine stock solution in PBS or 100 µl of PBS was added to each dish, and the dishes were returned to the incubator. The medium with either 4-thiouridine or PBS was exchanged every 2 h. 4 h after the start of the exposure, KSOM medium was removed, and 1 ml of TRIsure was added to the embryos to extract RNA. The RNA extraction method is described in the following section.

2.2 Molecular biology methods

2.2.1 Mouse genotyping

DNA from mouse ear-clips was isolated using the Sample-to-SNP kit (Life Technologies), and the DNA products were amplified using a ViiA7 real-time PCR machine (Life Technologies) with TaqMan assay kit (Life Technologies) and oligonucleotides listed in Table 2.2. Genotyping was performed with the support of the Sanger Institute RSF.

Table 2.2 Oligonucleotides used in this thesis

Target gene	Sequence ($5' > 3'$)	Assay type
UPRT-forward	ATTCCAAGATCTGTGGCGTC	TaqMan
UPRT-reverse	CTTCTCGTAGATCAGCTTAGGC	TaqMan
UPRT-probe (VIC)	CCGCATCGGGAAAATCCTCATCCA	TaqMan
Cre-forward	ACGTACTGACGGTGGGAGAA	TaqMan
Cre-reverse	GTGCTAACCAGCGTTTTCGTT	TaqMan
Cre-probe (VIC)	CTGCCAATATGGATTAACA	TaqMan
UPRT-forward	CCCGATATTCGACAAACGAC	SYBR Green
UPRT-reverse	GCTTCATGAGCACCACATTG	SYBR Green
Hprt-forward	GCCTAAGATGAGCGCAAGTTG	SYBR Green
Hprt-reverse	TACTAGGCAGATGGCCACAGG	SYBR Green

2.2.2 *In vitro* RNA synthesis

In vitro RNA synthesis reaction was performed with MAXIscript T7 Kit (Thermo Fisher) following the manufacturer's instructions. RNA Century-Plus Marker Templates (Thermo Fisher) was used as a DNA template. Two separate reactions were performed with and without 4-thio-UTP (Jena Bioscience). For the reaction with 4-thio-UTP, UTP and 4-thio-UTP were mixed at a 6:4 ratio. The reaction mix tubes were incubated at 37°C for 1 h and

were DNase-treated for 15 min to digest the template DNA. RNA was precipitated by adding 1/10 volume of 3 M sodium acetate, 20 µg of glycogen, and 3 volumes of 100% ethanol, and incubated at -20°C overnight. The resulting pellet was washed with 80% ethanol and resuspended with nuclease-free water.

2.2.3 Biotinylation of thiolated RNA

MTSEA biotin-XX (Biotium) was dissolved in *N,N*-dimethylformamide (DMF; Sigma-Aldrich) to prepare 0.1 mg/ml solution. Reaction mix was prepared by mixing 10 µl MTSEA biotin-XX solution, 15 µl 10X biotinylation buffer (100 mM HEPES, 10 mM EDTA; pH 7.5), and 105 µl nuclease-free water. 10 µl RNA was added to the reaction mix and incubated at room temperature for 2 h on a rotator with aluminium foil for protection from light. The RNA was cleaned up with the phenol:chloroform isolation method and was dissolved in nuclease-free water.

2.2.4 Isolation of biotinylated RNA with streptavidin beads

Biotinylated RNA was isolated with μMACS Streptavidin Starting Kit (Miltenyl Biotech). 25 µl of the biotinylated RNA was mixed with 100 µl μMACS streptavidin beads and incubated with rotation at room temperature for 15 min with aluminium foil. μColumns were placed in the magnetic field of the μMACS separator and equilibrated with 2 x 100 µl nucleic acid equilibration buffer supplied with the beads. The mixture of RNA and beads was applied to the equilibrated columns, and the flow-through was collected in a microfuge tube. μCoumns were washed twice with 500 µl of high salt wash buffer (100 mM Tris-HCl, 10 mM EDTA, 1 M NaCl, 0.1% Tween-20). RNA bound to the beads was recovered with 2 x 100 µl of 100 mM DTT. RNA in the flow-through and the eluent fractions was recovered with the phenol:chloroform isolation method.

2.2.5 RNA extraction from murine tissues

RNA was extracted with an acid guanidinium thiocyanate-phenol-chloroform extraction method (Chomczynski and Sacchi, 1987) with either TRIsure (Bioline) or TRIzol LS (Thermo Fisher) following the manufacturer's instructions. Tissue homogenisation was performed with different methods optimised for each tissue type.

Solid tissues

A piece of tissue was taken out of RNAlater, and residual RNAlater was removed from the sample with a Kimwipe (Kimberly-Clerk). The tissue was cut into a smaller piece that was around 30 mg in weight and placed in a 2-ml tube. A 7-mm stainless steel bead (Qiagen) and 1-ml TRIsure (Bioline) were added on top of it, and homogenisation with a TissueLyser LT (Qiagen) was performed until no visible debris was observed (2-3 min).

Serum

750 µl of TRIzol LS (Thermo Fisher) was added to 250 µl of serum and was pipetted up and down 10 times to ensure homogenisation.

Sperm

1 ml of TRIsure was added to a sperm pellet, and the solution was passed through a 25G needle attached to a 1-ml syringe multiple times until the solution became homogeneous, and no visible pellet remained.

After homogenisation, the solution was left still at room temperature for 3 min. 200 µl of chloroform-isoamyl alcohol mix (Sigma-Aldrich) was added, and the solution was mixed vigorously for 15 s followed by 3 min incubation at room temperature. The mixture was then centrifuged at 12,000g for 15 min at 4°C. The upper aqueous phase was transferred to a new 1.5-ml tube. To perform RNA precipitation without oxidising the thiol group, 1 µl 100 mM dithiothreitol (DTT; Promega), 1 µl 20 mg/ml glycogen (Thermo Fisher), and 500 µl 2-propanol (VWR) were added to the solution, and the mixture was stored at -20°C overnight after vigorous mixing with a vortex. RNA pellet obtained by centrifugation was washed with 80% ethanol and was resuspended in nuclease-free water.

2.2.6 RNA quantification and quality check

The concentration of RNA was quantified with a fluorescence-based quantification method specific to RNA using either Qubit RNA BR Assay kit or Qubit RNA HS Assay kit (Thermo Fisher), depending on the expected RNA concentration of samples. Quantification was performed following the manufacturer's instructions, and fluorescence was detected with Qubit fluorometer 2.0 (Thermo Fisher).

To confirm integrity of the RNA isolated from the tissues, the length distribution of the isolated RNA was assayed with Bioanalyzer (Agilent). It performs electrophoresis on a small chip, and the lengths of RNA are determined based on the migration pattern. Since rRNA is

known to show a specific pattern in degraded RNA samples, a RNA integrity number (RIN) is automatically determined based on the pattern of rRNA peaks by Bioanalyzer software. A value of 1 to 10 is assigned, with 10 being the least degraded. Below, a representative RNA profile of a sample with a high RIN is shown (Fig. 2.2).

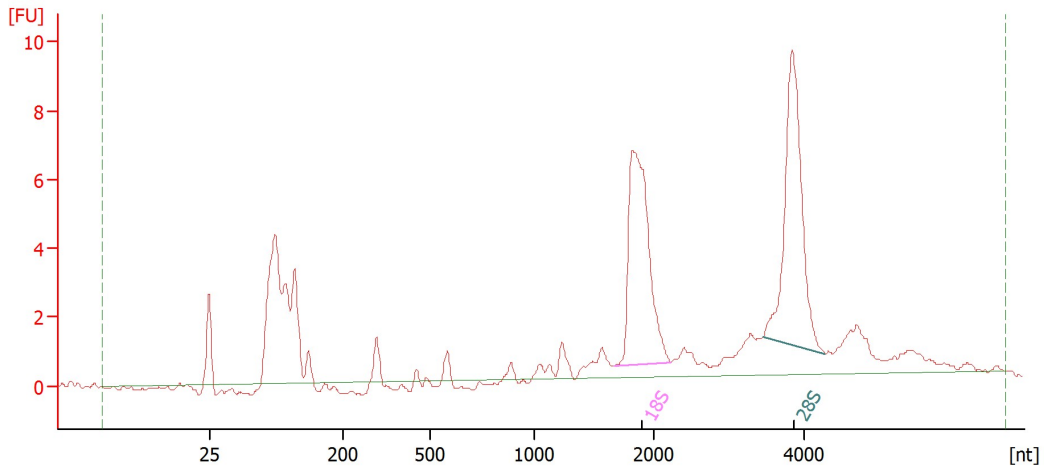


Fig. 2.2 Representative Bioanalyzer profile of high-quality RNA
Bioanalyzer software detects rRNA peaks based on RNA length profile. The rRNA peaks (shown as "18S" and "28S") are used to calculate an RIN. This sample has a RIN of 8.9, and the electropherogram shows a typical high-quality RNA profile. The 25-nt peak corresponds to the lower marker included in the assay buffer; FU, fluorescence unit.

2.2.7 RT-qPCR

Residual DNA was removed from RNA samples by a deoxyribonuclease (DNase) reaction using TURBO DNA-free kit (Thermo Fisher) following the manufacturer's instructions. DNase-treated RNA was then cleaned up with RNA Clean & Concentrator-5 (Zymo Research) following the manufacturer's instructions.

The abundance of UPRT and Hprt cDNA was quantified with PowerUp SYBR Green reagent (Thermo Fisher) and the primers listed in Table 2.2. Real-time polymerase chain reaction (PCR) was performed with the "Fast" mode on StepOnePlus real-time PCR machine (Thermo Fisher), which consists of an initial denaturation step at 95°C for 20 s, and then 40 cycles of 95°C for 3 s, followed by 60°C for 30 s for amplification and detection. Expression of UPRT in different samples was compared by the ΔΔCt method using Hprt as an internal control.

2.2.8 SLAMseq

Alkylation reaction was performed by preparing a 50-µl reaction mix (5-10 µg RNA, 10 mM IAA, 50 mM pH8 sodium phosphate, and 50% DMSO) followed by incubation at 50°C for 15 min. The reaction was stopped by adding 1 µl 1 M DTT. To precipitate the alkylated RNA, 1 µl glycogen (20 mg/ml), 5 µl NaOAc (3M, pH 5.2; Thermo Fisher), and 125 µl 100% ethanol were added, and the solution was incubated at -20°C overnight prior to RNA precipitation.

The alkylated RNA was used as an input for RNA sequencing library preparation. For polyA RNA-seq, QuantSeq 3′ mRNA-Seq Library Prep Kit for Illumina (Lexogen) was used for the library preparation. After confirming library profile with Bioanalyzer high sensitivity DNA assay kit (Agilent) (Fig. 2.3), a single-end 100 bp run was performed on Illumina HiSeq 1500.

For small RNA-seq, NEXTflex Small RNA-seq Kit V.3 (Bioo Scientific) was used for library preparation. PCR-amplified libraries were loaded on a 6% TBE (tris-borate-EDTA) gel (Thermo Fisher) and electrophorased at 200 V for 30 min. The gel was submerged in SYBR gold (Thermo Fisher) dissolved in TE buffer for 20 min to stain the DNA. Using a non-UV transilluminator, pieces of the gel containing DNA bands corresponding to libraries between 130-200 bp in length were cut out, and the libraries were recovered from the gel following the manufacturer's instructions. After confirming length profile of the resulting libraries with Bioanalyzer high sensitivity DNA analysis kit (Agilent) (Fig. 2.3), a single-end 50 bp run was performed on Illumina HiSeq 1500.

For embryo RNA analysis, RNA extracted from approximately 40-60 embryos per biological replicate was used for each library construction. TruSeq Stranded Total RNA Library Prep Gold Kit (Illumina) was used to capture rRNA-depleted total RNA. After confirming library profile with Bioanalyzer high sensitivity DNA assay kit (Agilent), a single-end 100 bp run was performed on Illumina HiSeq 1500.

2.2.9 Confirmation of alkylation reaction with a spectrophotometer

To confirm that the alkylation reaction was properly performed, a control reaction with 4-thiouracil was performed in parallel, using the same reagents and experimental conditions. A shift in absorption peak of 4-thiouracil solution was measured as a readout of successful alkylation. 1 µl of 10 mM 4-thiouracil solution was added to a 9-µl reaction mix (50% DMSO, 1 µl 100 mM IAA, 50 mM sodium phosphate; all final concentration), and the mix was incubated at 50°C for 15 min. After the reaction, the solution was diluted in 1:10, and 2 µl of it was used for UV-Vis analysis on NanoDrop spectrophotometer.

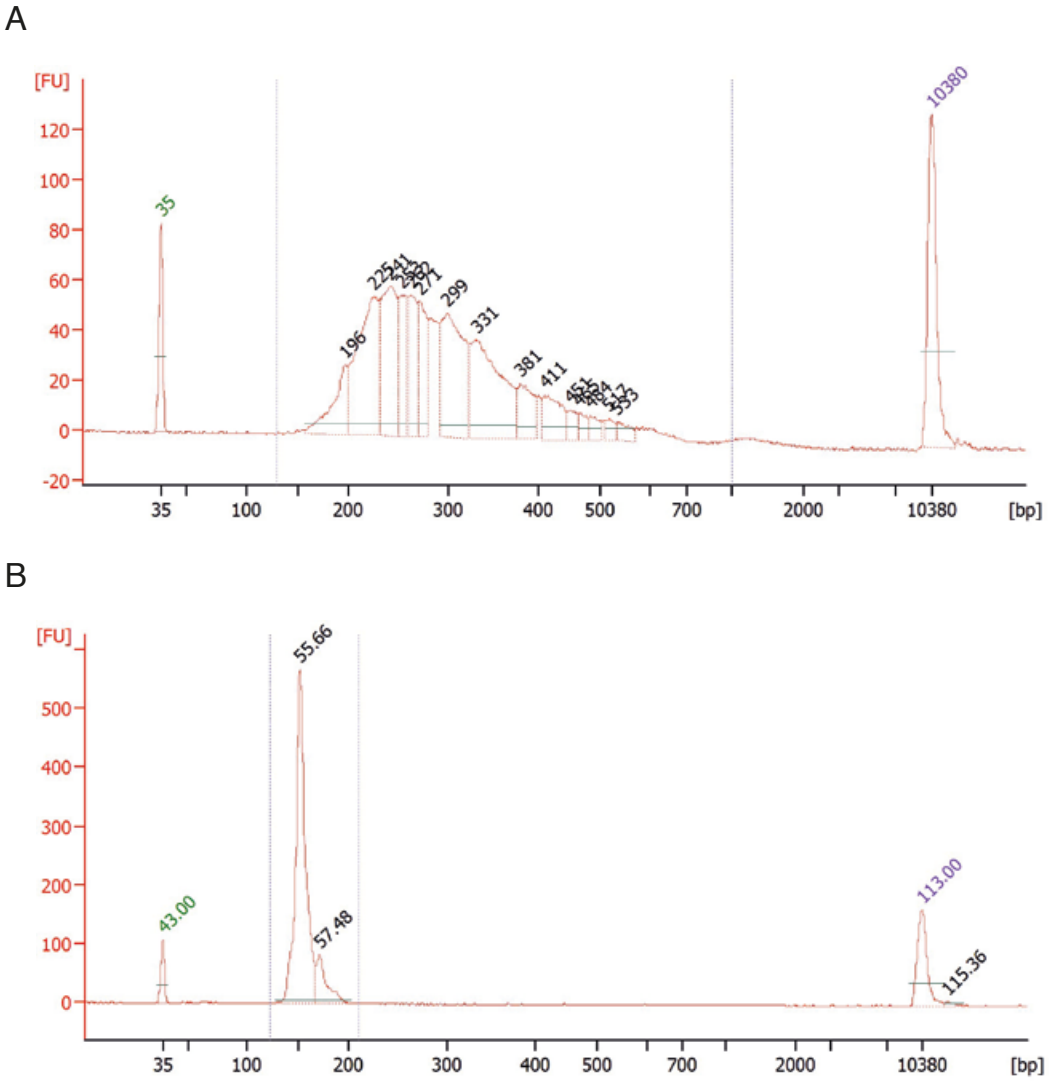


Fig. 2.3 Representative Bioanalyzer profile of successful RNA-seq libraries
(A) Library prepared with QuantSeq. A single peak is observed in the 200-300 bp region. (B) Library prepared with NEXTflex. A single sharp peak was observed in the 150-200 bp region. Both libraries do not show visible peaks at either the 100 bp or the >1000 bp regions, suggesting that little adapter dimers and PCR bubbles were generated. The 35- and 10380-bp peaks correspond to the lower and upper markers included in the assay reagent, respectively. FU, fluorescence unit.

2.3 Computational methods

2.3.1 Quality check of high-throughput sequencing

For all the RNA-seq runs, the demutiplexed reads were first analysed with FastQC[1] for a quality check. To ensure that a successful sequencing run was performed, it was confirmed that high base-calling quality was achieved along reads. The number of reads obtained for the libraries and the sequence quality are summarised in Appendix A. Also, libraries were checked to confirm that no overrepresented sequences are present, which may happen as a result of contamination.

2.3.2 SLAMseq analyses

PolyA RNA-seq

The reads were aligned to the *M. musculus* primary genome assembly (GRCm38) obtained from ensembl[2] by SLAM-DUNK (v0.3.3) (Herzog et al., 2017; Neumann et al., 2019) with options `"-t 5 -5 12 -n 100 -m -mv 0.2 -c 2 -rl 100"` (see Table 2.3 and SLAM-DUNK website[3] for further details). 3′ UTR annotation was constructed by merging all the 3′ UTR regions of Refseq genes obtained from UCSC Table Browser[4] (assembly: GRCm38) and Ensembl genes obtained from BioMart[5] (Ensembl Genes 87). This 3′ UTR data was used to resolve the ambiguity of multimapping reads and to measure read count and T>C rate. This analysis outputs T>C conversion rate detected for each gene, and this table was further analysed to identify genes with significantly higher T>C rate in Cre^+ mice compared with Cre^- mice using an R package `ibb` (Pham et al., 2010). Procedures and R script used to perform ibb analysis are shown in Appendix B.

Small RNA-seq

The adapter sequence (`TGGAATTCTCGGGTGCCAAGG`) was removed from the demultiplexed reads with cutadapt software (Martin, 2011). Since RNAs that are too short cannot be assigned to a single genomic locus, and reads originated from small RNA (20-30 nt) should contain the 3′ adapter sequence in a 50 bp sequencing run, reads shorter than 20 bp after trimming and reads containing no adapter sequence were removed (`"-m 20 --discard-untrimmed"`).

[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[2] http://www.ensembl.org/
[3] http://t-neumann.github.io/slamdunk/docs.html
[4] https://genome.ucsc.edu/cgi-bin/hgTables
[5] https://www.ensembl.org/biomart/martview

Table 2.3 Description of arguments used in SLAM-DUNK

Argument	Description
`-t`	Number of threads to use
`-e`	Use an end-to-end alignment algorithm for mapping
`-5`	Length (bp) to remove from the 5′ end of all reads
`-a`	Maximum length (bp) of polyA allowed at the 3′ end of reads
`-n`	Maximum number of alignments to report per read
`-m`	Use reference to resolve multimappers
`-mv`	Minimum variant fraction to call SNPs
`-mi`	Minimum alignment identity to retain reads at the filter step
`-c`	Minimum coverage to call SNPs
`-rl`	Maximum read length (bp) in the supplied BAM file

Since the 5′ and 3′ adapters included in NEXTflex kit have 4 random bases at their ligating end, the first and last 4 bases were further removed from the adapter-trimmed reads using cutadapt with "`-u 4 -u -4`" options. The processed reads were aligned to the *M. musculus* genome (GRCm38) by SLAM-DUNK (v0.3.3) with options "`-mv 0.2 -rl 50 -a 50 -5 0 -mi 1 -e -m -n 5000`". Labelled transcripts were discovered with ibb software as described in Appendix B.

Total RNA-seq

The adapter sequence (`AGATCGGAAGAGCACACGTCTGAACTCCAGT`) was removed from the demultiplexed reads with cutadapt software. The processed reads were aligned to the *M. musculus* genome (GRCm38) by SLAM-DUNK (v0.3.3) with options "`-mv 0.2 -rl 50 -5 0 -mq 0`". TE annotation was obtained from Repbase[6] (v20140131). Labelled transcripts were discovered with ibb software as described in Appendix B. For labelled TE discovery, duplicate-level T>C count and read count were first summarised at gene level, and then an ibb analysis was performed.

2.3.3 Reanalysis of the published FACS dataset

To construct lists of genes that are expressed in endothelial cells and other brain cells in mice, the data table containing FPKM (fragments per kilobase of transcript per million mapped reads) value for each gene in multiple cell types isolated from mouse brain was used (Zhang et al., 2014). FPKM >0.1 was used as a conservative threshold to determine

[6]https://www.girinst.org/

the expressed genes (>99% confidence). Also, in this dataset, FPKM values smaller than 0.1 were rounded up to 0.1 to avoid inflated values when calculating ratios. Thus, the genes detected in the sorted endothelial cells were determined by filtering genes that have FPKM >0.1 in endothelial cells; the genes expressed only in non-endothelial cells were identified by choosing the genes with a mean FPKM >0.1 among non-endothelial cells and FPKM = 0.1 in endothelial cells.

2.3.4 Gene ontology (GO) term enrichment analysis

The list of all the annotated genes was sorted by *P*-values obtained from the beta-binomial test in ascending order and then used as input for PANTHER 13.1[7] (Mi et al., 2017) to perform the "Statistical enrichment test" with full biological processes GO terms. The enriched GO terms obtained were further analysed using REVIGO[8] (Supek et al., 2011) with allowed similarity = 0.4 to better visualise the results with less redundancy of GO terms.

2.3.5 Motif enrichment analysis

HOMER[9] (Heinz et al., 2010) was used to discover the enrichment of known TF-binding motifs upstream of labelled genes in the 2-cell embryos. The script `findMotif.pl` was run on the list of labelled genes selected at FDR <0.1 using the default parameters.

[7] http://pantherdb.org/
[8] http://revigo.irb.hr/
[9] http://homer.ucsd.edu/homer/

Chapter 3

Development of *in vivo* cell type-specific RNA labelling

3.1 Background

Multicellular organisms are supported by a number of organs that further consist of numerous highly specialised cells, which have distinct gene regulatory networks. Thus, it is essential to accurately monitor gene expression in each type of specialised cells to better understand the functions of each tissue and animal physiology.

Cell-type-specific gene expression analysis has commonly been achieved through isolation of cells of interest prior to either a low- or high-throughput gene expression assay. Laser capture microdissection (LCM) was developed to dissect cells of interest out of a cryosectioned tissue through a microscope (Emmert-Buck et al., 1996). Although the notable benefit with this method is that the cells can be histologically oriented, maintaining the intact tissue structure, LCM is relatively low-throughput and was shown to be noise-prone compared to other cell isolation methods (Okaty et al., 2011). This is because it is not always possible to distinguish different types of cells through their appearances, and contamination from neighbouring cells can occur.

One of the most commonly employed methods to isolate cells is fluorescence-activated cell sorting (FACS) (Julius et al., 1972). Cells of interest are labelled through either by generating transgenic animals expressing a fluorescent marker under a cell type-specific promoter or by tagging cells with fluorescently-labelled antibody against a cell-specific antigen. The cells are dissociated into single cell level and are sorted by a flow cytometer, which not only measures fluorescent intensity, but also estimates the cell diameter to better achieve isolation of a particular cell population. Although FACS is a very robust method and

has been applied in wide-ranged research fields in biology, one of the major drawbacks of the method is that it requires dissociation of an animal/tissue so that singlet cells can be obtained. It sometimes needs extensive optimisation to achieve optimal dissociation maintaining cell viability, and, even after such optimisations, this isolation step itself could potentially affect the gene expression of the cells (Richardson et al., 2015).

Recent advance in RNA-seq methods even enables to sequence RNA obtained from single cells (scRNA-seq: single-cell RNA-seq) (Tang et al., 2009). Dimensionality reduction analyses, combined with prior knowledge about marker genes, enable studying transcriptome of each single cell without sorting them. This approach is also promising in identifying previously-unknown cell types or subpopulations within a known cell type by finding cells that do not cluster with other known cell types. However, there are a few reports stating that the difference in tissue dissociation methods affect the transcriptome of the cells, and some unique cell clusters discovered with scRNA-seq could be due to artefacts induced from a tissue dissociation method employed (O'Flanagan et al., 2019; van den Brink et al., 2017).

Alternative approaches to study cell-type-specific RNA that do not depend on cell isolation have also been developed. Instead of isolating cells prior to an analysis, molecules are "tagged" in a cell-type-specific manner, and the tagged molecules are assayed after isolating the molecules from the entire tissue (Fig. 3.1). One of such method, TU tagging, is reviewed below.

3.1.1 TU tagging

TU tagging uses a uracil analogue, 4-thiouracil, to label RNA in a cell-type-specific manner (Cleary et al., 2005; Gay et al., 2013). Uracil phosphoribosyltransferase (UPRT) from *Toxoplasma gondii* (*T. gondii*) converts 4-thiouracil to 4-thiouridine monophosphate (4-thio-UMP), which is further converted to 4-thiouridine triphosphate (4-thio-UTP) to be incorporated into the newly synthesised RNA. Although the mammalian genome encodes a *UPRT* gene, it was shown to have little enzymatic activity, and thus wild-type mammalian cells cannot salvage 4-thiouracil to incorporate it into newly synthesised RNA (Cleary et al., 2005; Li et al., 2007). By generating mice expressing *T. gondii* UPRT in a specific type of cells, cell-type-specific RNA labelling can be achieved through systemic 4-thiouracil administration. For the remainder of this thesis, "UPRT" refers to *T. gondii* UPRT unless otherwise stated.

To achieve cell-type-specific UPRT expression in mice, Gay et al. developed a transgene that expresses UPRT in a Cre recombinase (Cre)-inducible manner. The transgene consists of a chicken beta-actin (CA) promoter followed by a green fluorescent protein (GFP)-coding cassette, which is flanked by the locus of X-over P1 (loxP) sequences, and a UPRT-coding

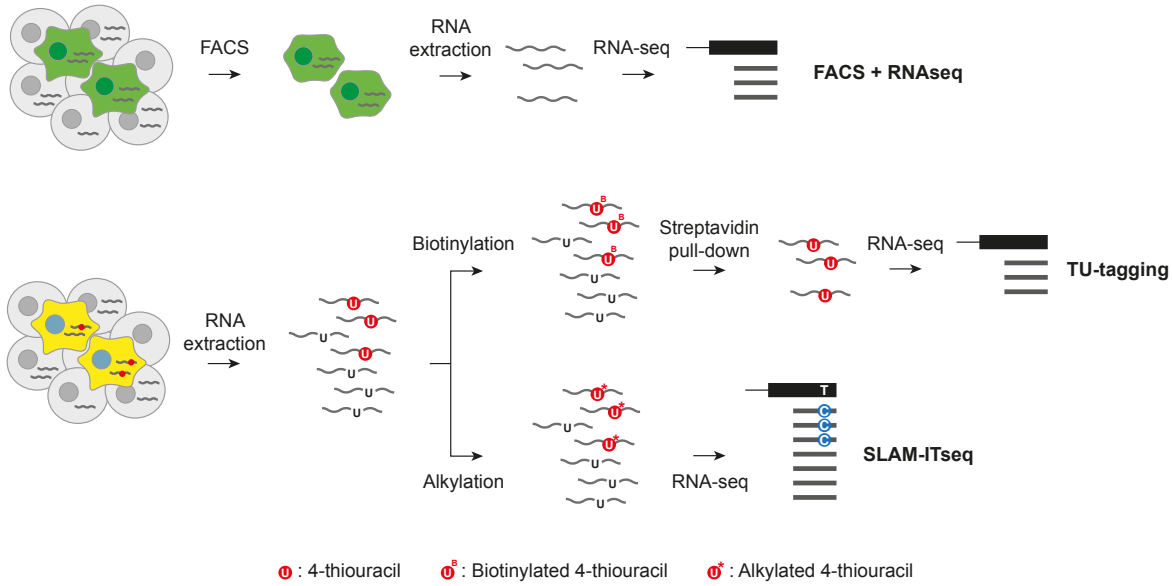


Fig. 3.1 Comparison of different cell-type-specific transcriptomics methods
Schematics of experimental approaches of FACS, TU tagging, and SLAM-ITseq are shown (taken from Matsushima et al. (2018)). With FACS, cells of interest are physically isolated, and RNA-seq is performed on RNA extracted from the isolated cells. On the other hand, the other two methods metabolically label RNA in a cell-type-specific manner, and the RNA is extracted from the entire tissue/animal of interest without cell sorting. TU tagging identifies the labelled RNA through thiol-specific biotinylation followed by streptavidin pull-down, while SLAM-ITseq identifies the labelled U through thiol-specific alkylation, which introduces T>C base conversions at the thiolated uridine positions.

sequence (Fig. 3.6). Since the GFP cassette includes SV40 polyadenylation sequences, this transgene only expresses GFP where no Cre is expressed, whereas, in Cre-expressing cells, the GFP cassette is removed, resulting in UPRT expression. Thus, by crossing mice with this UPRT transgene (*UPRT* mice) with mice bearing Cre transgene under a cell-type-specific promoter (*Cre* mice), double-transgenic mice that express UPRT only in the Cre-expressing cells are obtained.

To detect the 4-thiouracil-labelled RNA, biotinylation of the thiol group followed by streptavidin pull down is performed to first isolate the labelled RNA. The abundance of RNA in the pulled-down fraction as well as the input RNA is measured by RNA-seq, and the thiolated RNA is identified by finding RNAs that are enriched in the pulled-down fraction (Fig. 3.1).

Although this TU tagging has been applied in a few research projects (Chatzi et al., 2016; Erickson and Nicolson, 2015; Gay et al., 2013), this method has some limitations mostly due to its use of a biochemical method to isolate the labelled RNA.

First, the biochemical isolation methods of RNA are noise- and bias-prone (Duffy et al., 2015). The bead-based isolation methods have some false-positive signals that are due to the unspecific binding of RNA to the beads. RNA with longer lengths were shown to be preferentially pulled down than shorter ones. Also, potentially due to its repetitive washing step as well as inefficient elution step, its recovery rate is not always high, which may lead to lower sensitivity. Although an improved biotinylation method has been developed (Duffy et al., 2015), and it improved the pull-down efficiency, it still does not eliminate the need of an isolation step with beads.

Second, identification of the labelled RNA from a comparison between the eluted and input fractions is challenging. Since only newly synthesised transcripts are labelled and pulled-down, the compositions of each RNA species in the pulled-down and input fractions are completely different, which makes normalisation as well as identification of enriched transcripts challenging. Most frequently used software for differential gene expression analysis use statistical models that assume more or less similar proportions of each RNA species across samples for normalisation (Anders and Huber, 2010; Robinson et al., 2010), and thus these software are not appropriate to be employed for TU tagging analyses. One solution to overcome this normalisation problem would be to include spike-in RNA in each fraction; however, optimisation of the ratio between spike-in and total RNA is not always straightforward. Too much or too little spike-in concentration in each library may lead to over- or under-representation of spike-in-derived reads, respectively. Since it is not easy to predict how much RNA is labelled in the tissue before performing an experiment, the spike-in concentration must always be optimised for each experiment. Even though an appropriate

amount of spike-in is used, unbiased discovery of actively transcribed genes might not be achievable. Since TU tagging only measures the enrichment level of each transcript in the eluent fraction compared to the flow-through fraction, a threshold of the enrichment level to determine which transcripts are enriched needs to be set. Although the selection of such a threshold greatly affects the results, there has not been enough data to conclude which threshold yields the best sensitivity and specificity.

3.2 Aims of this chapter

To further improve the TU tagging method, I developed a new method, SLAMseq in tissue (SLAM-ITseq), which combined the TU tagging with SLAMseq. Also, I redesigned the mating strategy to obtain the transgenic animals in order to control for non-specific RNA labelling. With these changes, the labelled RNA can be identified with little bias.

First, to analyse the recovery rate and detection bias with the conventional pull-down method used in TU tagging, *in vitro* transcribed labelled RNA was used. Pull-down experiments on the synthetic labelled RNAs revealed the detection bias introduced by the method.

To test the performance of SLAM-ITseq in comparison to TU tagging, *in vivo* RNA labelling was performed on endothelial cells in the mouse brain. Since the same transgenic animals were also previously used for TU tagging experiments, direct comparison of the sensitivity and specificity of these two methods were accomplished.

Next, to test its robustness, SLAM-ITseq was applied to two other tissues using different transgenic strains. The labelled gene lists in each cell type validated the specificity and sensitivity of the method.

3.3 Assessment of the pull-down method with *in vitro* synthesised RNA

To assess how well the pull-down based method recovers thiolated RNA, thiol-RNA synthesised *in vitro* was used as input for the pull-down method, and the recovery efficiency was assayed. DNA Template mix with different lengths (100, 200, 300, 400, 500, 750, and 1,000 nt) was used to synthesise RNA with corresponding lengths, and *in vitro* RNA transcription was performed with and without 4-thio-UTP (at the ratio of UTP:4-thio-UTP = 6:4) (Fig. 3.2). The synthesised RNA was biotinylated with methanethiosulfonate ethylammonium-biotin (MTSEA-biotin), which was shown to be an efficient thiol-specific biotinylating agent (Duffy

et al., 2015). The biotinylated RNA was then isolated with streptavidin beads, and both eluent and flow-through fractions were recovered. The abundance as well as length distribution of the RNA were assayed with Bioanalyzer.

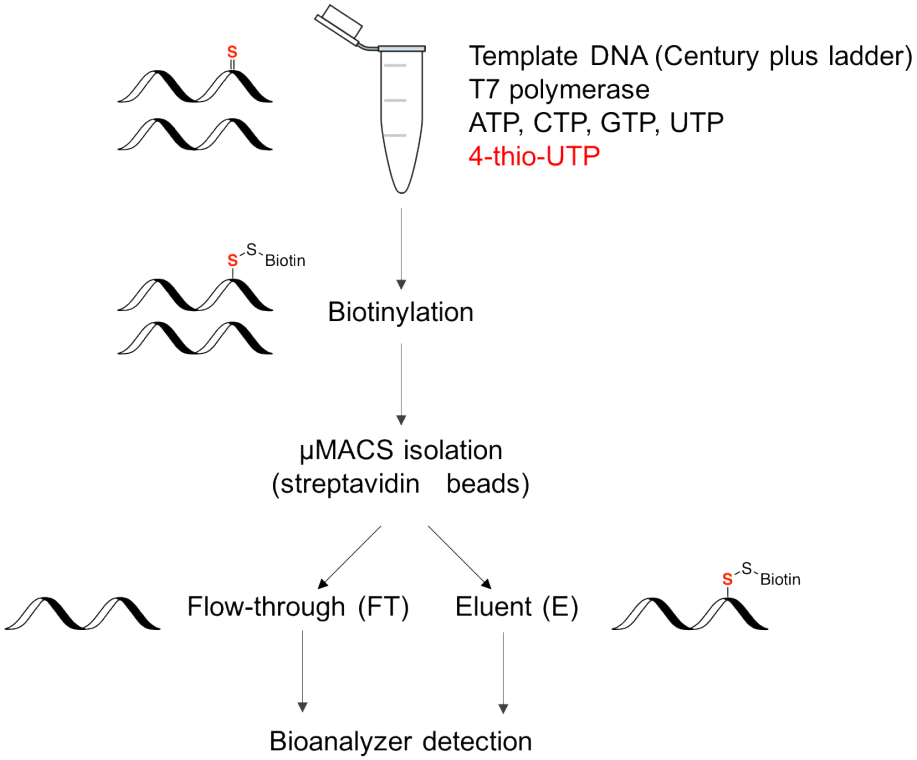


Fig. 3.2 Schematic of the pull-down assay with *in vitro* synthesised RNA
In vitro synthesised RNA with and without 4-thio-UTP was biotinylated, and then the labelled RNA was isolated with streptavidin beads. RNA bound to the beads was eluted and collected as the eluent fraction (E), and RNA that did not bind to the beads was also recovered as the flow-through fraction (FT). Both fractions were assayed with Bionalayzer for length and abundance profiling.

Bioanalyzer results show that little RNA is obtained in the eluent fraction from RNA synthesised without 4-thio-UTP in the eluent, which suggests that this method has a low false-positive rate (Fig. 3.3, 3.4). On the other hand, only around 30% of the labelled RNA was recovered in the eluent fraction, and 15% of the input appeared in the flow-through fraction.

Since the labelled RNA was synthesised with the presence of normal UTP, not all the RNA synthesised contains 4-thio-UTP. To better measure the recovery rate of the labelled RNA, the RNA in the eluent fraction, which should consist solely of the labelled RNA, was used for another round of biotinylation and streptavidin isolation steps.

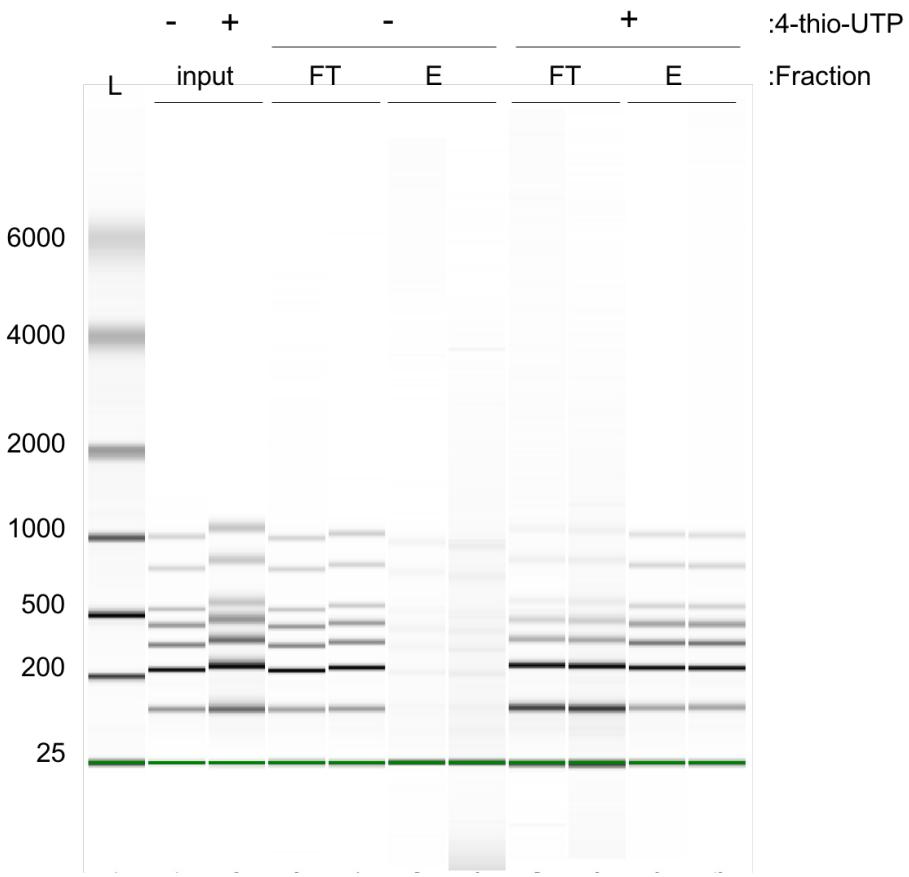


Fig. 3.3 Bioanalyzer results of the pull-down assay
RNA profile obtained with Bioanalyzer is shown for the flow-through (FT) and eluent (E) fractions of RNA synthesised with and without 4-thio-UTP. Note that the band intensity is adjusted for each sample, and thus cannot be used to compare the relative abundance of RNA among samples. L, ladder in nt.

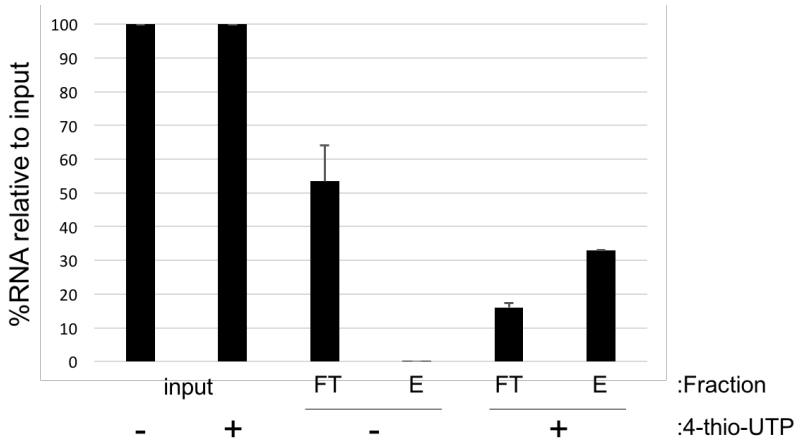


Fig. 3.4 Abundance of RNA obtained from different pull-down fractions
Relative abundance of RNA in the flow-through (FT) and eluent (E) fractions obtained from the RNA synthesised with and without 4-thio-UTP is shown. Error bars represent standard deviations among technical replicates ($N = 2$ each).

The recovery rate of the labelled RNA was revealed to be high (80-%) (Fig. 3.5B), and only around 10% of the input was not captured and lost in the flow-through fraction. However, the Bioanalyzer profile shows that the RNA that was lost in the flow-through fraction was biased towards RNA with shorter lengths (Fig. 3.5A).

These results suggest that although thiol-RNA-specific recovery can be achieved with the biotin-streptavidin isolation method, this approach may favour longer RNAs, potentially because they contain more 4-thio-UTPs per molecule.

3.4 Development of SLAM-ITseq

3.4.1 Design of SLAM-ITseq

To better achieve cell-type-specific RNA labelling *in vivo*, I have redesigned TU tagging and adapted SLAMseq to detect labelled RNA. For simplicity, I name this new method as SLAM-ITseq (SLAMseq in tissue) (Matsushima et al., 2018, 2019) (Fig. 3.6).

There are a few potential sources of false-positive signal with the use of 4-thiouracil and UPRT transgene for cell-type-specific analysis. First, although mammalian UPRT has been reported to be inactive, this claim was made based on a chemoluminescence assay (Cleary et al., 2005), which may not be the most sensitive method, and thus mammalian UPRT could potentially have a very minimal enzymatic activity. Also, mammalian cells may have another metabolic pathway to convert 4-thiouracil to 4-thio-UTP. Second, although 3

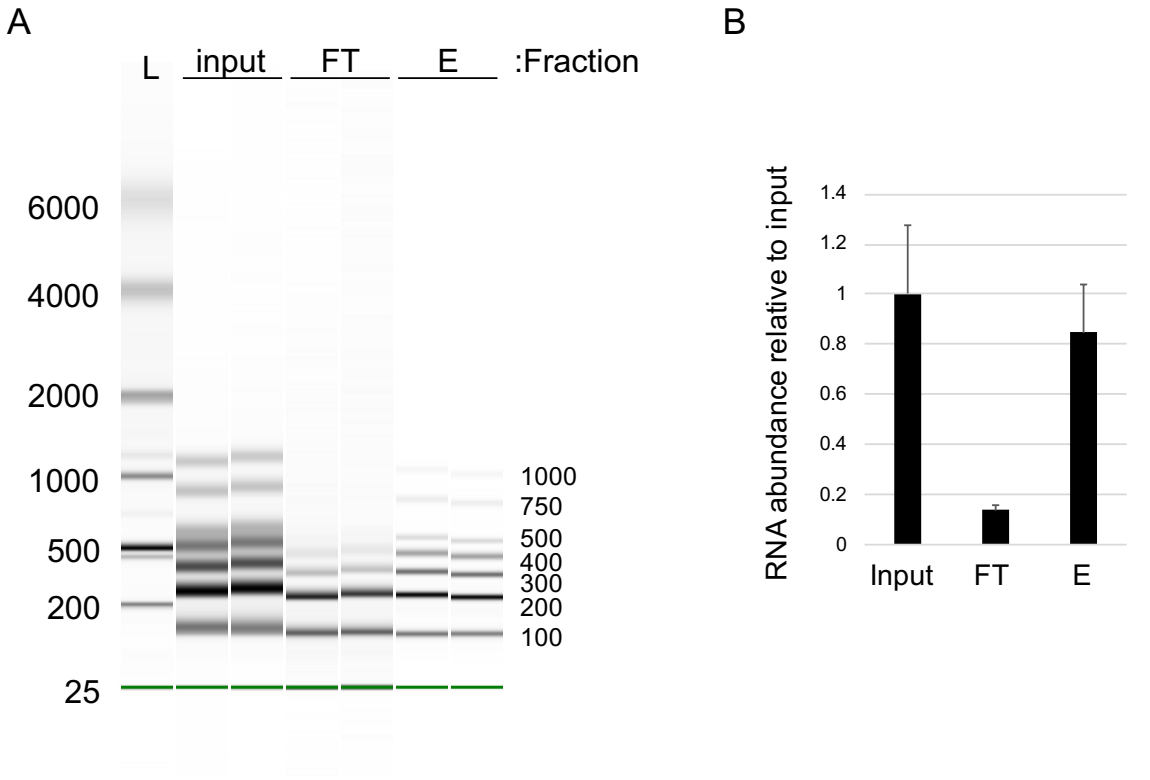


Fig. 3.5 Recovery rate of labelled RNA
The pulled-down labelled RNA was treated with another round of an isolation step to assay the length profile and concentration of recovered RNA. (A) Bioanalyzer results showing the RNA profile of the flow-through (FT) and eluent (E) fractions. Band intensity is adjusted for each sample and thus cannot be compared among samples. L, ladder in nt. (B) Relative abundance of RNA obtained in FT and E is summarised as a bar graph. Error bars represent standard errors among technical replicates (N = 2 each).

consecutive SV40 polyadenylation sequences are placed between GFP-coding sequence and UPRT-coding sequence for an efficient termination of transcription, there may be a low-level read-though of the termination signals by Pol II.

To control for the UPRT-independent 4-thiouracil incorporation and Cre-independent UPRT expression, I decided to include a control mice that only have the UPRT transgene. To obtain this transgenic strain with maximum genetic background similarity, homozygous *UPRT* mice (*uprt/uprt*) and hemizygous *Cre* mice (*cre/0*) were crossed so the double transgenic mice (*uprt/0*; *cre/0*) and hemizygous UPRT mice (*uprt/0*; +/+) were obtained in a 50:50 ratio in F1 (Fig. 3.6A). For simplicity, I refer the double transgenic mice as Cre^+ and the hemizygous *UPRT* mice as Cre^-. To identify the significantly labelled transcripts in Cre^+ mice compared to Cre^- mice, the beta-binomial regression was employed to test if the observed difference of conversion rates between Cre+ and Cre- is significant based on variance among biological replicates in each genotype (Pham et al., 2010).

3.4.2 Confirmation of Cre-inducible UPRT expression

To confirm if SLAM-ITseq can be applied for a cell-type-specific transcriptome analysis, I used *Tie2:Cre* mice, which express Cre only in endothelial cells (Kisanuki et al., 2001), to generate Cre^+ mice. The same Cre strain was also used in TU tagging paper, so the sensitivity and specificity of these two methods can be compared directly (Gay et al., 2013). Cre^+ and Cre^- mice generated as described in the previous section received intraperitoneal (i.p.) 4-thiouracil injection and, the tissues were collected 4 h after the injection.

First, to confirm if Cre-inducible UPRT expression was achieved, RNA from an entire mouse brain was collected, and reverse transcription followed by quantitative polymerase chain reaction (RT-qPCR) against UPRT was performed (Fig. 3.7). Significantly higher UPRT expression was observed in Cre^+ brain compared to Cre^-, which suggests that UPRT was expressed in a Cre-inducible manner.

3.4.3 Quality-check of the alkylating reaction

Since it is known that 4-thiouracil with and without a carboxyamidomethyl group has different absorption spectra, ultraviolet-visible (UV-Vis) spectrophotometry was performed on 4-thiouracil after an alkylation reaction to validate that the alkylation reaction was successfully performed (Fig. 3.8). The absorption spectra show that an absorption peak was observed around 330-340 nm in control 4-thiouracil solution, while the peak is shifted to 300 nm in the 4-thiouracil reacted with IAA. This result indicates that successful alkylation was performed using the reagent and the incubation condition used.

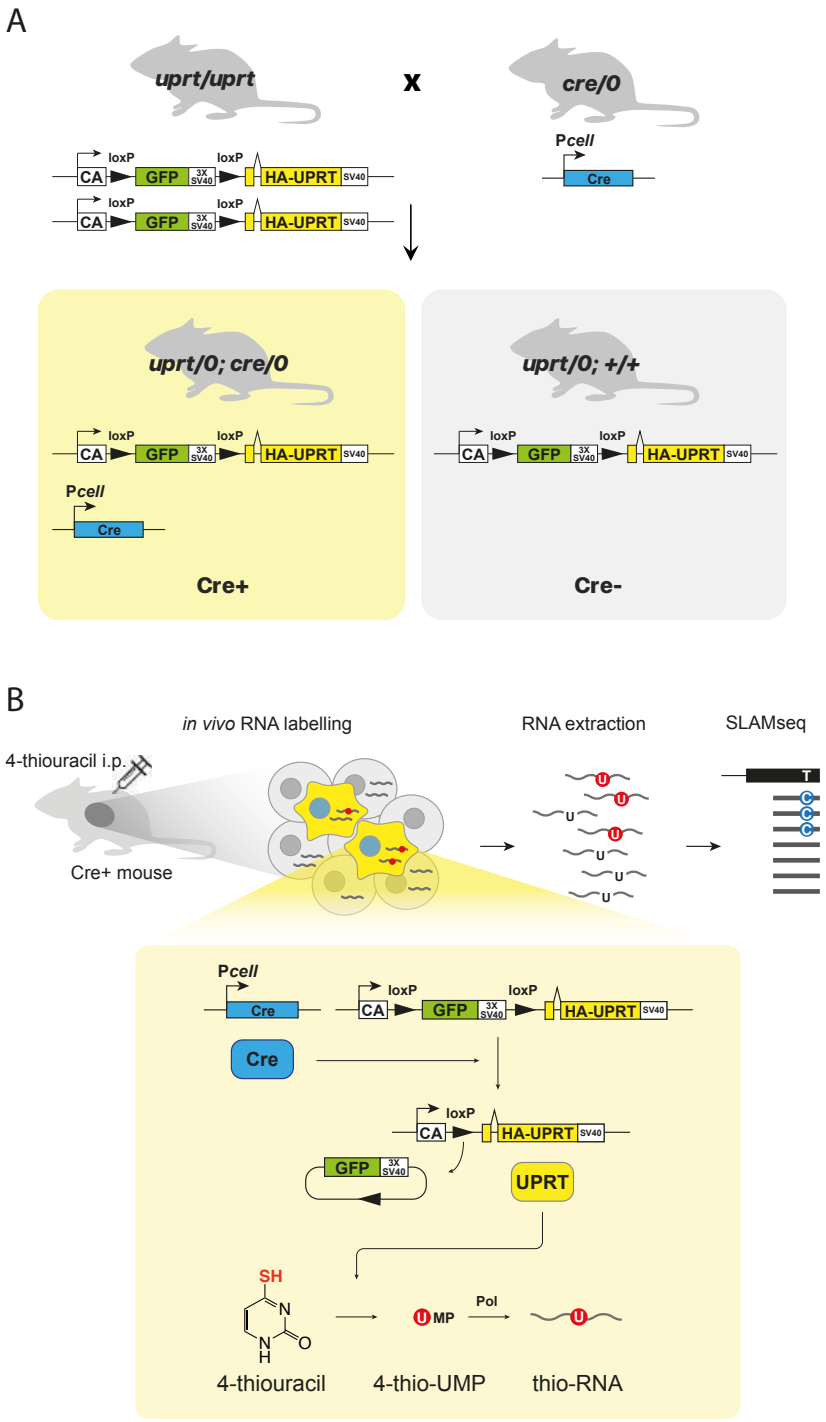


Fig. 3.6 Schematic representation of the SLAM-ITseq design
(A) Cross design for a SLAM-ITseq experiment is shown. Cre^+ and Cre^- mice are generated by crossing homozygous *UPRT* mice and hemizygous *Cre* mice. (B) Both Cre^+ and Cre^- mice are exposed to 4-thiouracil intraperitoneally. In the Cre-expressing cells, the GFP-coding cassette is removed by Cre, resulting in UPRT expression. 4-thiouracil is incorporated into newly-synthesised RNA only in the Cre-expressing cells. RNA extracted from a tissue that contains the Cre-expressing cells is analysed with SLAMseq, and RNA synthesised in the Cre-expressing cells is identified by finding higher T>C mismatches induced by 4-thiouracil incorporations. Taken from Matsushima et al. (2018).

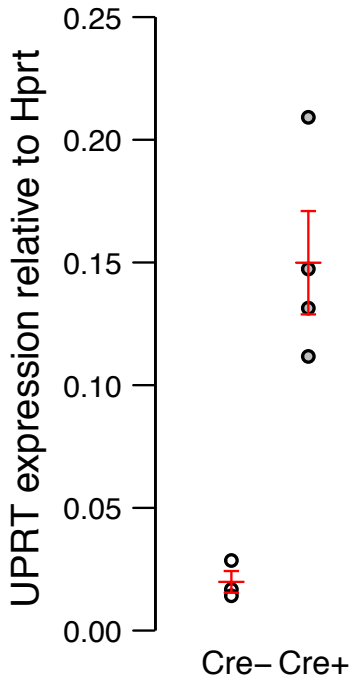


Fig. 3.7 UPRT expression in Cre$^+$ and Cre$^-$ mice
qRT-PCR was performed against UPRT in the brains of Cre$^+$ and Cre$^-$ mice. Relative expression of UPRT to Hprt is plotted, and each point represents a biological replicate (N = 3 for Cre$^-$ and N = 4 for Cre$^+$). Red horizontal lines represent the mean and 95% confidence intervals. Taken from Matsushima et al. (2018).

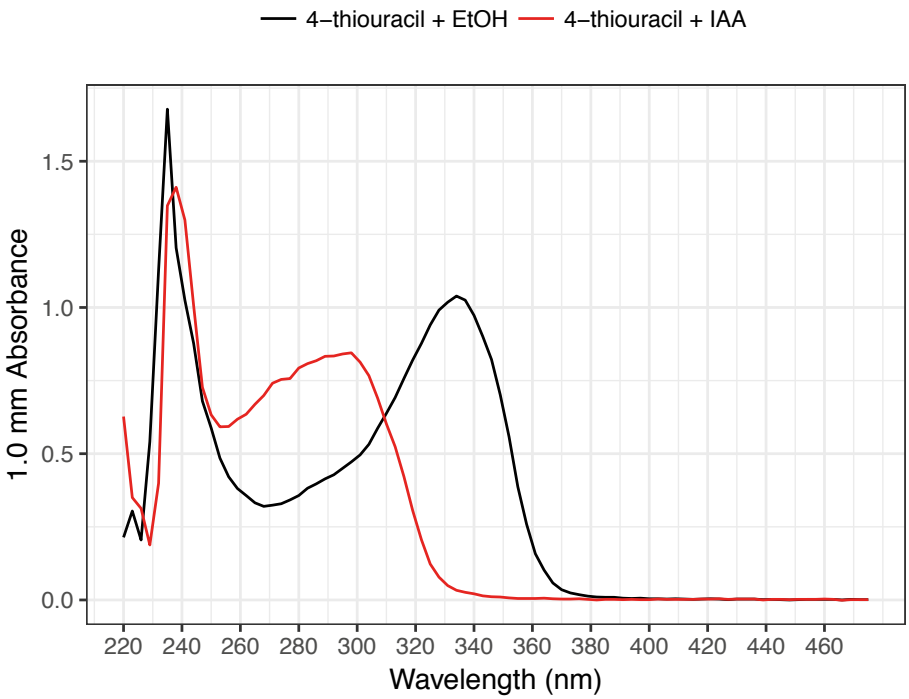


Fig. 3.8 Absorption spectra of 4-thiouracil with and without IAA treatment
UV-Vis spectrophotometry was performed on 4-thiouracil solution with and without the alkylation reaction with IAA. Taken from Matsushima et al. (2019).

3.4.4 Higher overall labelling level was observed in Cre$^+$ brain

RNA isolated from the brains of Cre$^+$ and Cre$^-$ mice were treated with IAA and used to prepare RNA-seq library. High-throughput sequencing was performed on the IAA-treated RNA, and the obtained reads were analysed with SLAM-DUNK software, which was designed for SLAMseq analyses (Herzog et al., 2017; Neumann et al., 2019).

Next, to confirm successful 4-thiouracil incorporation in Cre$^+$ mice, T>C rate in the two strains were compared (Fig. 3.9). As expected, significantly higher T>C conversions were observed in Cre$^+$ brain compared to Cre$^-$ brain.

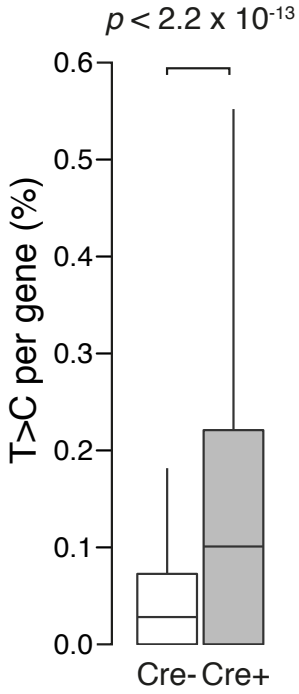


Fig. 3.9 T>C rate comparison between Tie2-Cre$^+$ and Tie2-Cre$^-$ brain
Distributions of T>C rate for each gene in Cre$^+$ and Cre$^-$ are shown as boxplots (taken from Matsushima et al. (2018)). The lower and upper hinges correspond to the first and third quartiles, the middle hinges indicate the median, and the whiskers extend to 1.5 interquartile range from the upper hinges. Outliers are not shown. Two-tailed Mann–Whitney *U*-test was used to calculate the *P*-value indicated.

3.4.5 Labelled transcripts were identified by beta-binomial test

SLAMseq followed by SLAM-DUNK analysis outputs the number of Ts sequenced and T>C conversions for each transcript, and for each biological replicate. To best identify the

transcripts with a higher T>C rate in Cre^+ compared to Cre^- considering variance among biological replicates, the beta-binomial regression was employed. In each sample, events of T>C conversions follow the binomial distribution, and the ratio of T>Cs to all Ts sequenced follow the beta distribution among the replicates. This model has been used to identify differentially methylated regions from bisulfite sequencing data (Dolzhenko and Smith, 2014; Feng et al., 2014; Sun et al., 2014), where non-methylated Cs are converted to Gs after the bisulfite treatment, while methylated Cs are protected from the conversion.

Using the beta-binomial model, the probability density function Pr can be shown as below:

$$Pr(M = m|n, \alpha, \beta) = \binom{n}{m} \frac{B(m+\alpha, n-m+\beta)}{B(\alpha, \beta)}$$

where the number of 4-thio-UTPs incorporated m and the total number of Ts sequenced n. B is the beta function and $p \sim B(\alpha, \beta)$ (shape parameters $\alpha \geq 0,\ \beta \geq 0$) is assumed.

The beta-binomial test was performed using an R package, ibb (Pham et al., 2010), and, at the threshold of Benjamini-Hochberg FDR (false discovery rate) <0.05, 5,427 genes were significantly labelled (Fig. 3.10A). Also, to confirm if the presence of Cre or UPRT in the tissue induced any aberrant transcription, the abundance of RNA between Cre^+ and Cre^- was compared (Fig. 3.10B). The abundance of RNA in the two genotypes is highly and positively correlated, suggesting that the presence and effect of the Cre transgene has little effect on the transcriptome of the tissue.

Interestingly, the labelled gene list not only includes endothelial-specific genes, but also genes expressed globally, the so-called "house-keeping" genes (Fig. 3.11). This result is important in assessing the sensitivity of SLAM-ITseq. The proportion of endothelial cells in all the cells that consist mouse brain is thought to be approximately <5% (Gay et al., 2013). Since the house-keeping genes are expressed in all the cell types, the labelled house-keeping transcripts synthesised in endothelial cells are diluted in >95% of the unlabelled transcripts when SLAM-ITseq was performed on the mouse brain. This result suggests that SLAM-ITseq is sensitive enough to detect labelled RNA that consists less than 5% of the total pool of the particular transcript.

Since 4-thio-UTPs are incorporated at the positions of Ts in each gene, the number of Ts in each gene could potentially introduce a bias in identifying labelled transcripts. To assess this potential bias, the distributions of the number of Ts in the labelled and total transcripts are compared (Fig. 3.12). The plot revealed that the labelled transcripts tend to contain more Ts compared with all the transcripts detected. This could be due to the fact that transcripts with more Ts can incorporate the analogue with higher probability.

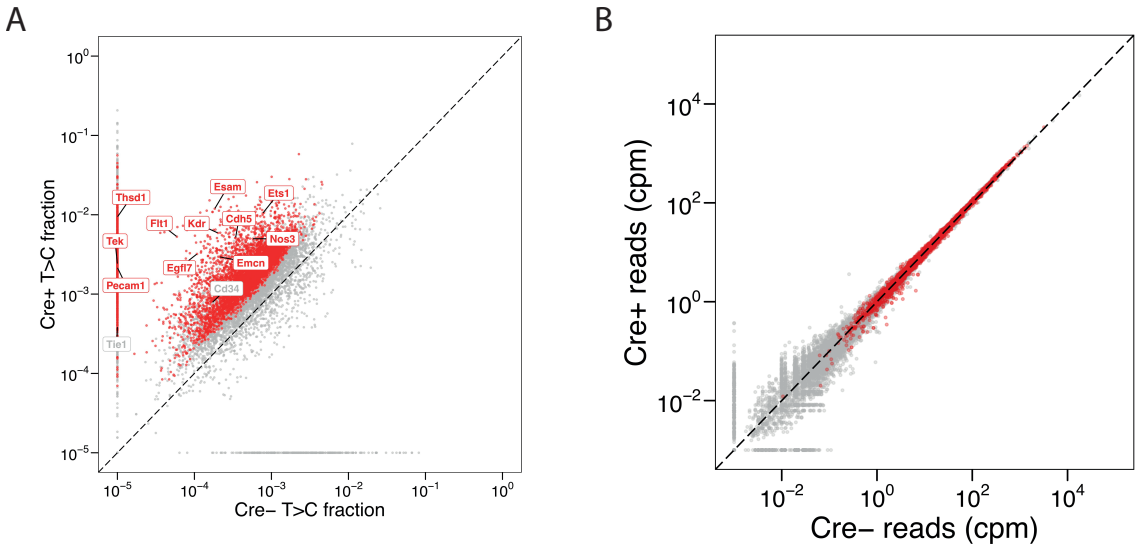


Fig. 3.10 Labelled genes identified by SLAM-ITseq
(A) T>C levels of RNA in Tie2-Cre^{+} and Tie2-Cre^{-} brain are compared. Each plot represents the mean value among the biological replicates in each genotype (N = 4 for Cre^{+} and N =3 for Cre^{-}). Red points show the significantly labelled (FDR <0.05, beta-binomial test) genes, and the known endothelial genes are marked with a label. A constant value of 10^{-5} is added to each value when plotting. (B) The abundance of each transcript is compared between the genotypes. A high correlation (Pearson's correlation coefficient = 0.99) was observed between the two. A constant value of 10^{-3} was added to each value when plotting. Taken from Matsushima et al. (2018).

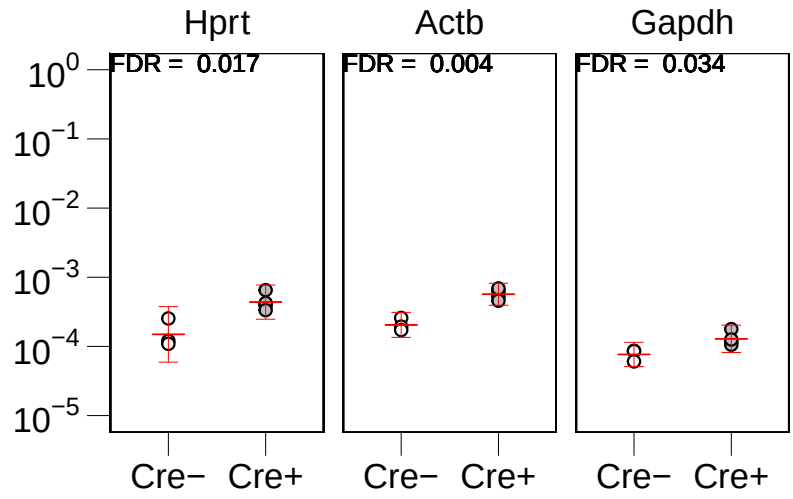


Fig. 3.11 Labelling levels of globally expressed genes
T>C conversion rates of three globally expressed genes are compared between Cre^{+} and Cre^{-}. Each point represents a value for a biological replicate, and red bars represent the mean and 95% confidence intervals among the replicates in each genotype. FDR (Benjamini-Hochberg procedure) obtained from the beta-binomial test is also shown. Taken from Matsushima et al. (2018).

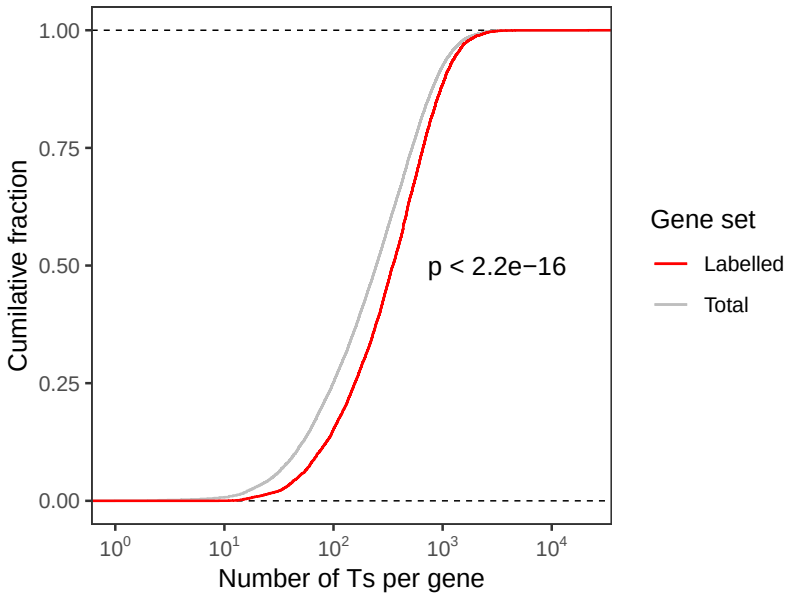


Fig. 3.12 Comparison of the number of Ts between labelled and total transcripts
Empirical cumulative distribution function plot summarises the distribution of the number of Ts in labelled transcripts and all the detected transcripts (total). Two-sided Kormogorov-Smirnov (KS) test was applied to assess the difference of the distributions, and the *P*-value from the test is shown.

To compare the sensitivity of SLAM-ITseq with that of TU tagging, a set of known endothelial genes that was used as the positive control in TU tagging was used (Gay et al., 2013). Among 13 known endothelial genes, 11 of them were labelled with SLAM-ITseq (Fig. 3.10A), which is comparable to the sensitivity of TU tagging. To perform a more comprehensive analysis on the sensitivity and specificity of SLAM-ITseq, a previously published dataset that identified the transcriptome of different types of cells in mouse brain by FACS was used (Zhang et al., 2014). Importantly, the endothelial cells in this study were fluorescently-labelled and isolated using a GFP transgene under the same promoter, *Tie2*. The SLAM-ITseq-labelled genes overlap with the FACS-identified endothelial genes, but its overlap with non-endothelial genes was less than 5%, suggesting that endothelial-specific RNA labelling was achieved with SLAM-ITseq (Fig. 3.13).

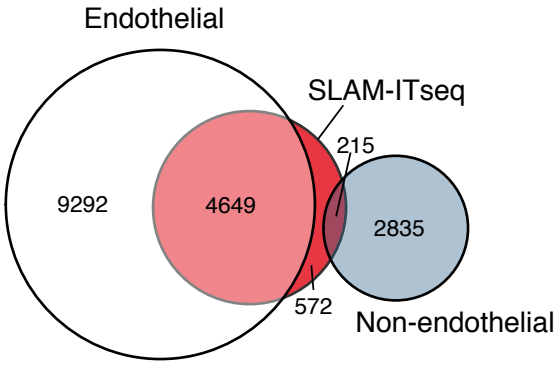


Fig. 3.13 Euler diagram comparing the genes identified with SLAM-ITseq and FACS The genes labelled in Tie2-Cre$^+$ mice by SLAM-ITseq (red circle) are compared with the genes identified by FACS (Zhang et al., 2014). The genes that were expressed in the FACS-sorted endothelial cells are shown as a white "Endothelial" circle, while the genes that were not detected in endothelial cells but expressed in other cell types in mouse brain are shown as a grey "Non-endothelial" circle. Taken from Matsushima et al. (2018).

Further, to comprehensively analyse what types of genes were labelled with SLAM-ITseq, gene ontology (GO) term enrichment analysis was performed on the labelled gene list (Fig. 3.14). GO terms that are known to be linked to endothelial functions, such as "cardiovascular system development", are enriched. This further confirms that endothelial-related genes are selectively labelled with SLAM-ITseq.

3.5 SLAM-ITseq application in two other murine tissues

To test the robustness of SLAM-ITseq, two other types of cells were additionally analysed: epithelial cells in intestine and adipocytes in white adipose tissue (WAT). For both cell

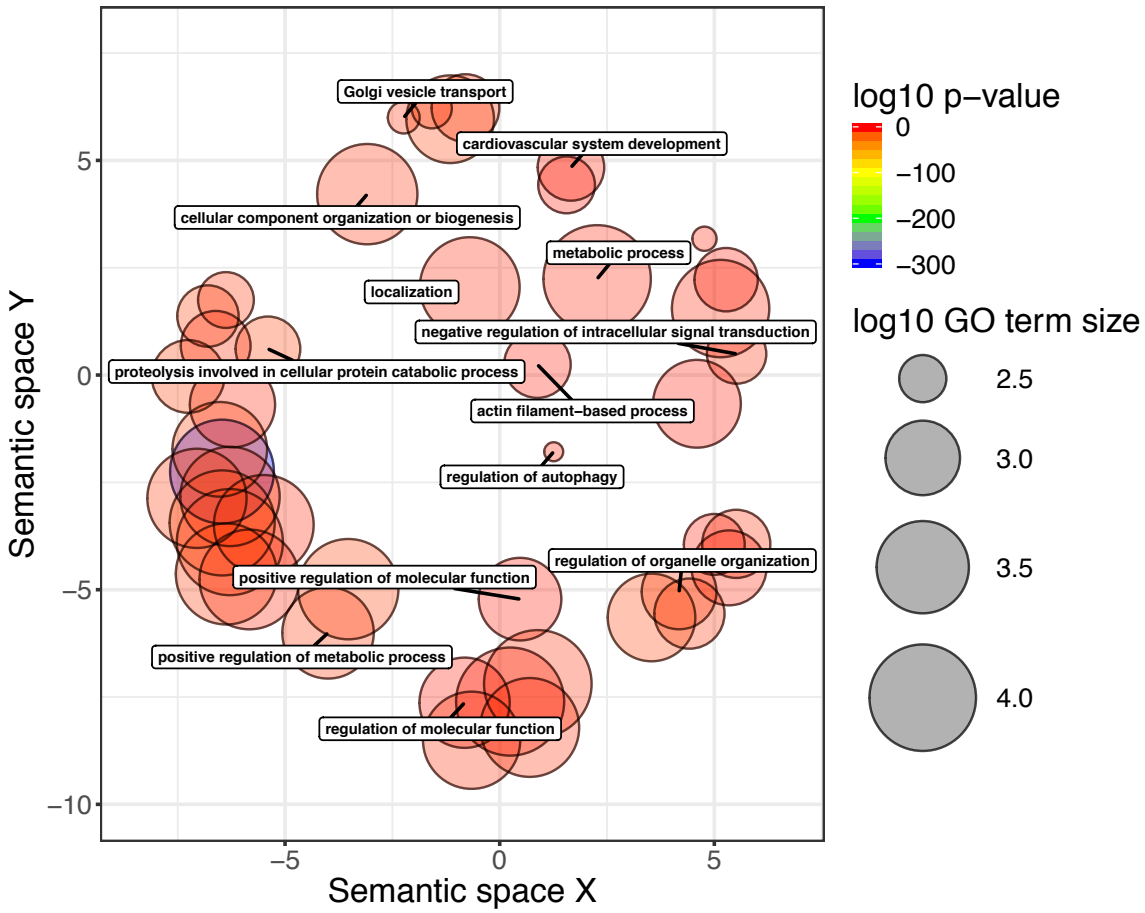


Fig. 3.14 GO term enrichment analysis performed on the labelled genes in Tie2-Cre$^+$
GO term enrichment analysis was performed on the labelled gene list obtained from Tie2-Cre mice using PANTHER (Mi et al., 2017), and similar GO terms were visualised as two-dimensional clusters, semantic scape X/Y, by REVIGO (Supek et al., 2011) (taken from Matsushima et al. (2018)).

types, well-characterised specific Cre lines are available: *Vil-Cre* (Madison et al., 2002) and *Adipoq-Cre* (Eguchi et al., 2011) (Fig. 3.17), and these lines were crossed with *UPRT* mice to generate double-transgenic lines.

3.5.1 UPRT expression was confirmed in Vil-Cre$^+$ and Adipoq-Cre$^+$

First, to confirm Cre-inducible UPRT expression was achieved in each tissue, RNA was extracted from duodenum of *Vil-Cre* mice and epididymal white adipose tissue (eWAT) of *Adipoq-Cre* mice and used for RT-qPCR analyses (Fig. 3.15). Significantly higher UPRT expression was observed in Cre$^+$ mice compared with Cre$^-$ mice in both strains, suggesting that UPRT is expressed in a Cre-inducible manner in these tissues.

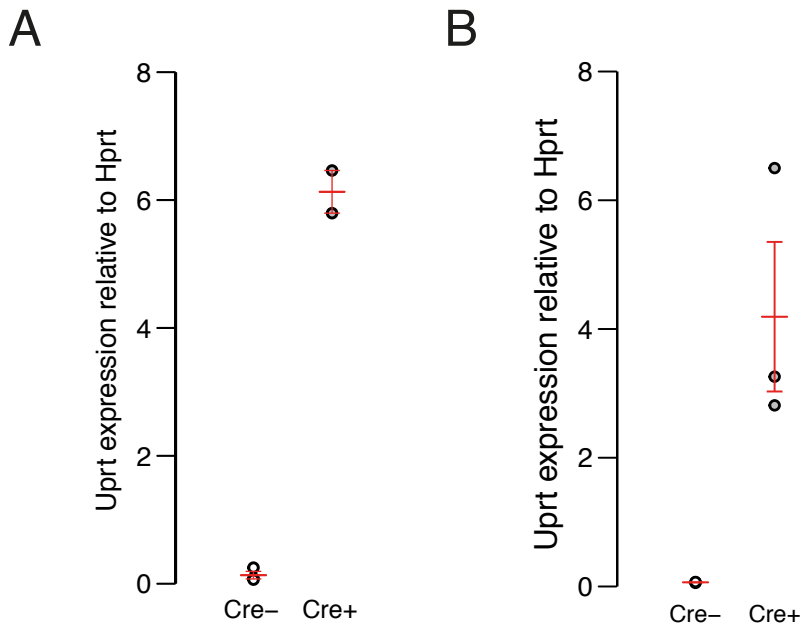


Fig. 3.15 UPRT expression analysis in Vil-Cre$^+$ and Adipoq-Cre$^+$ mice
RT-qPCR analysis was performed against UPRT in mice with (A) *Vil-Cre* and (B) *Adipoq-Cre*. Relative expression of UPRT to Hprt is quantified, and points representing a value for each biological replicate are shown. Red horizontal lines represent the mean of the values and 95% confidence intervals. Taken from Matsushima et al. (2018).

3.5.2 *In vivo* RNA labelling was achieved without transcriptome perturbation

Next, SLAMseq was performed on *Vil-Cre* mice and *Adipoq-Cre* mice, and an order of 1,000 genes with significantly higher T>C conversions were identified in each strain (Fig. 3.16).

Also, to assess if the UPRT or Cre expression had any effect on the transcriptome of the tissues assayed, the abundance of each transcript was compared between Cre^+ and Cre^- mice. In both strains, a high positive correlation between the genotypes was observed, suggesting that Cre or UPRT expression did not induce aberrant transcription.

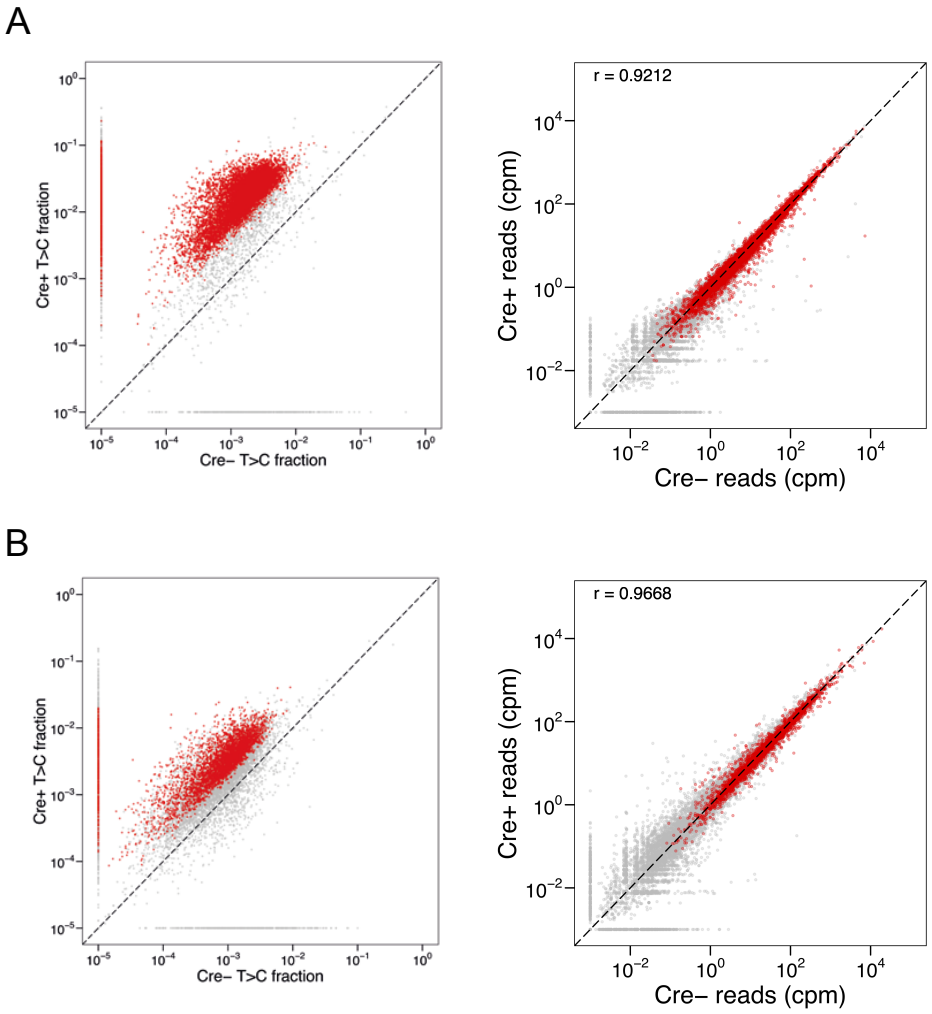


Fig. 3.16 T>C rate and abundance of each transcript in Cre^+ and Cre^-
T>C ratio (left panes) and RNA abundance (right panes) in Cre^+ and Cre^- are compared in (A) *Vil-Cre* and (B) *Adipoq-Cre*. Red points represent the genes with a significantly higher T>C rate (FDR <0.05) in Cre^+ mice compared to Cre^- mice. Each point shows the mean value among the biological replicates. A constant value of 10^{-5} and 10^{-3} were added to T>C fraction and read count (cpm), respectively. Pearson's correlation coefficient r is shown.

3.5.3 RNA labelling is specific to the Cre-expressing cells

Since both tissues from which RNA was extracted consist of heterogeneous cell types, the labelling levels of a few known marker genes for each cell type were compared to confirm the specificity of the labelling with SLAM-ITseq (Fig. 3.17A). While intestine consists of various types of cells such as epithelial cells, interstitial cells, smooth muscle cells, and endothelial cells, *Vil-Cre* is known to be only expressed in epithelial cells (Madison et al., 2002) (Fig. 3.17). In Vil-Cre^{+} mice, while known epithelial genes such as *Vil*, *Muc4*, and *Lyz1* were significantly labelled, non-epithelial genes such as *Kit* (interstitial), *Acta2* (smooth muscle), and *Pecam1* (endothelial) were not labelled.

WAT consists of adipocytes and endothelial cells. Some marker genes for each cell type were similarly tested in Adipoq-Cre^{+} mice to confirm the labelling specificity (Fig. 3.17B). Adipose marker genes (*Adipoq*, *Fabp4*, and *Pparg*) were confirmed to be significantly labelled, whereas endothelial genes (*Esam*, *Pecam1*, and *Thsd1*) were not labelled, suggesting that adipose-specific RNA labelling was achieved.

Second, to comprehensively assay what types of genes are labelled in each Cre line, GO term enrichment analysis was performed on the labelled gene lists (Fig. 3.18). In the Adipoq-Cre data (Fig. 3.18A), the analysis revealed that some adipose-related GO terms such as “fat cell differentiation” were enriched, while no enrichment of endothelial-related terms was found. This result suggests that the labelled gene list contains a high number of adipose-related genes. However, in the Vil-Cre^{+} data, no cluster linked to specific cellular functions was found (Fig. 3.18B). This may be due to the fact that Vil-Cre^{+} cells in intestine consist of various types of cells (e.g. enterocytes, Paneth cells, and intestinal stem cells), and thus no significant enrichment for particular GO terms was obtained.

These data suggest that SLAM-ITseq can achieve RNA labelling in wide-ranged Cre strains without labelling RNA in surrounding cells.

3.6 Discussion

In this chapter, the development of a novel *in vivo* RNA labelling method, SLAM-ITseq, by employing *UPRT* line and novel metabolic RNA labelling method, SLAMseq, was shown.

SLAM-ITseq utilises transgenes developed for TU tagging, but it is combined with a newly developed detection method, SLAMseq. Another major improvement to TU tagging was to use Cre^{-} animals to control for background labelling, which may be introduced by UPRT-independent 4-thiouracil incorporation, unfiltered SNPs, or sequencing errors. In fact, our data showed the presence of noticeable background labelling in Cre^{-} mice ranging from an order of 10^{-5} to 10^{-1}. Although it is unclear whether all the factors above are influential

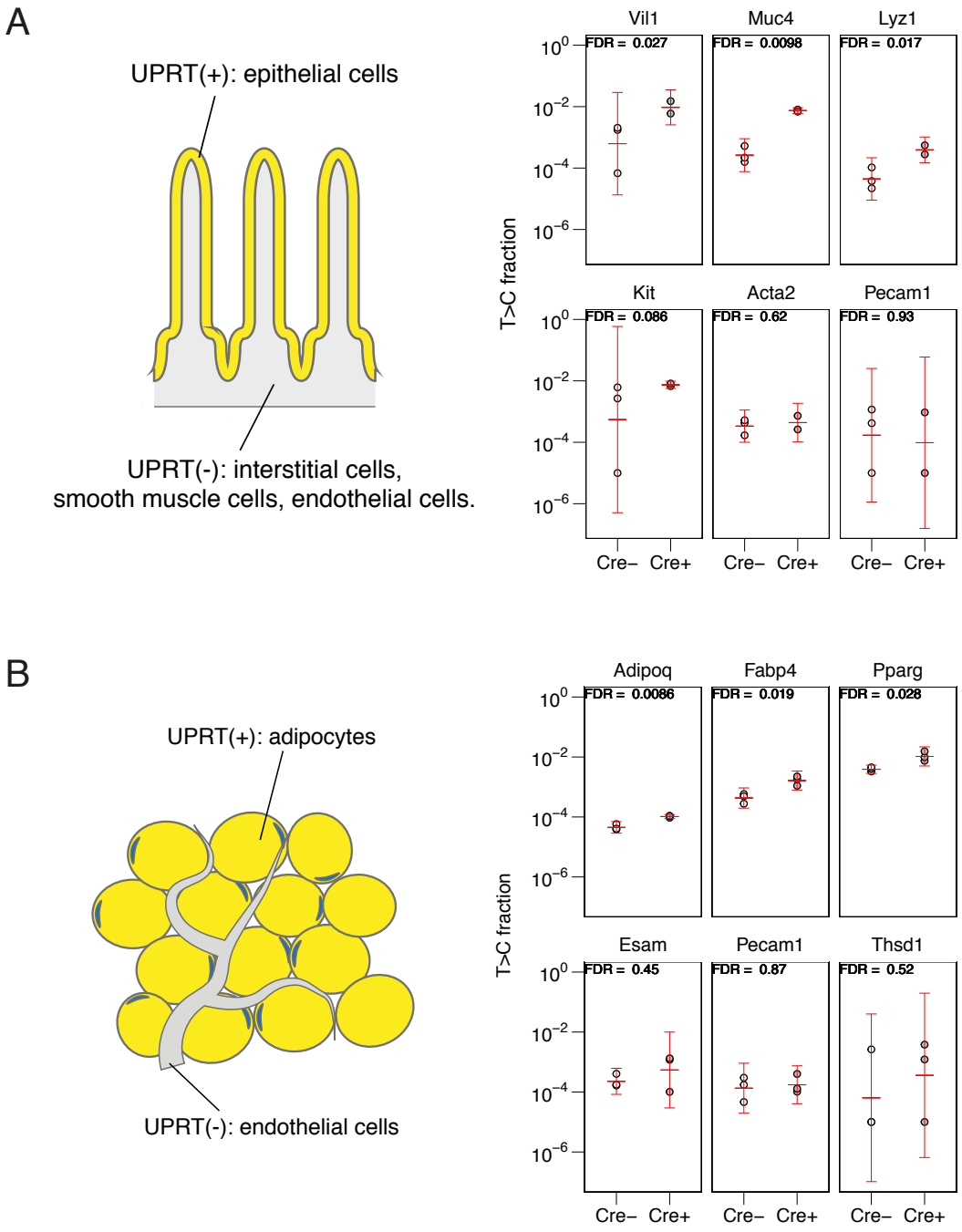


Fig. 3.17 T>C conversion rates of transcripts known to be expressed in Cre$^+$ and Cre$^-$ cells

A few marker genes for Cre$^+$ and Cre$^-$ cells in (A) Vil-Cre and (B) Adipoq-Cre were chosen to validate the specificity of SLAM-ITseq. On the left, histological schematics of the tissues that RNA was extracted from are shown, and the cells expressing Cre as well as UPRT are shown in yellow. On the right, genes shown in the top three panes are known marker genes for the Cre-expressing cells, while the genes in the bottom three panes are known to be expressed in the non-Cre-expressing cells of the tissues. Red horizontal bars represent mean T>C fraction and 95% confidence intervals among biological replicates. Benjamini-Hochberg FDR from the beta-binomial test is shown for each gene. Taken from Matsushima et al. (2018).

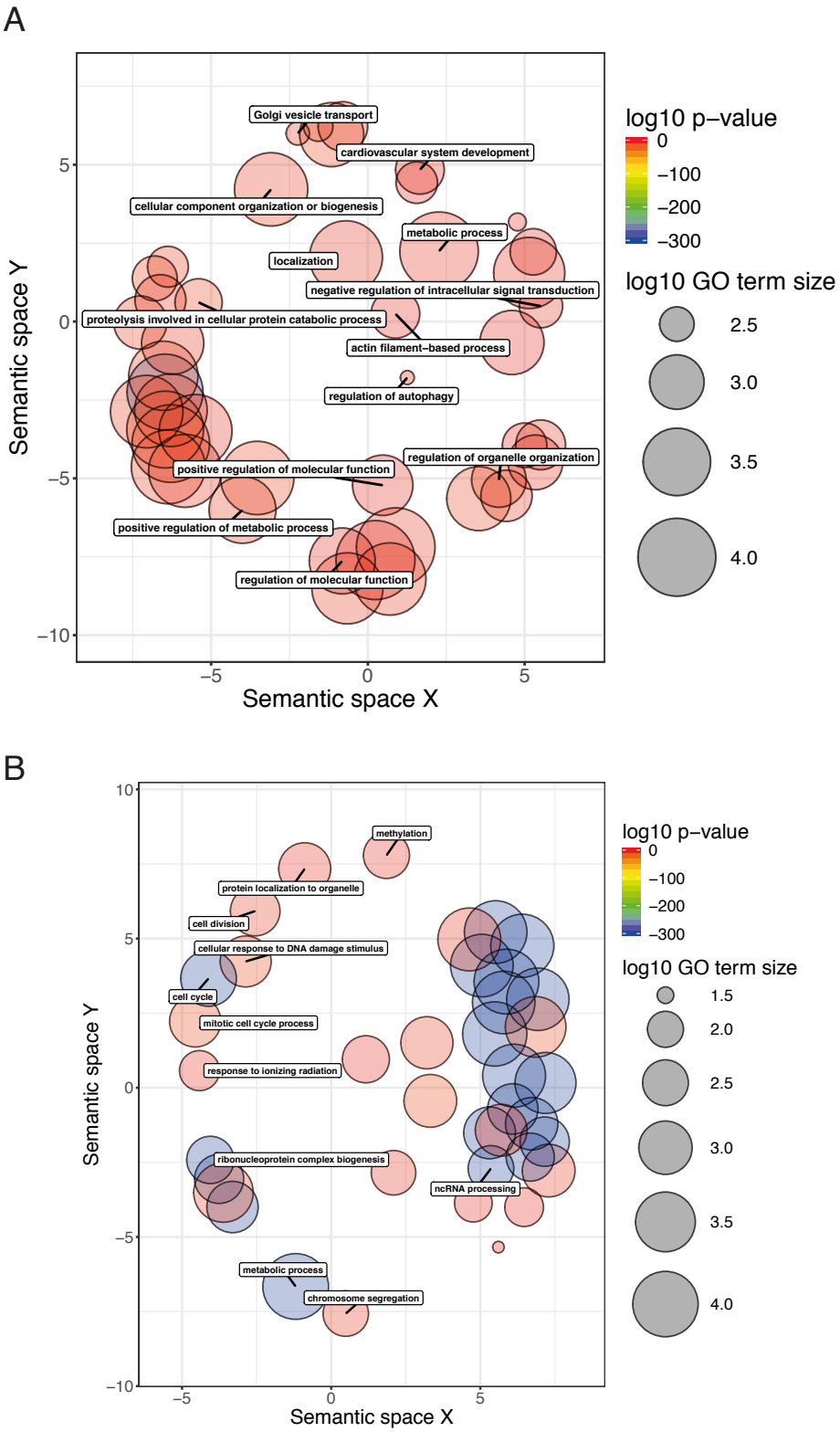


Fig. 3.18 GO term enrichment analysis on the labelled transcripts in Vil-Cre$^+$ and Adipoq-Cre$^+$

GO term enrichment analysis on the labelled gene list in (A) *Vil-Cre* and (B) *Adipoq-Cre* using PANTHER (Mi et al., 2017). Related GO terms are shown together as clusters using REVIGO (Supek et al., 2011). Taken from Matsushima et al. (2018).

or just one of them is critical, our data suggests the importance of using a control mice to control for this background labelling.

Cell-type-specific RNA labelling was confirmed to be achieved in the three different cell types: endothelial cells, intestinal epithelial cells, and adipocytes. It is important to note that ratios of Cre-expressing cells in these tissue are varied: endothelial cells account for only 5% of all cells in brain, whereas epithelial cells in duodenum and adipocytes in WAT comprise much higher proportions. Also, these three tissues exist in anatomically different locations of the mouse body, and thus these cells may have different availability of 4-thiouracil from the circulation. These facts suggest that SLAM-ITseq can achieve robust cell-type-specific RNA labelling regardless of a Cre-line used, a ratio of Cre-expressing to non-Cre-expressing cells, and anatomical locations of tissues analysed.

3.6.1 Comparison with other methods

SLAM-ITseq achieves cell-specific transcriptome analysis without physically isolating cells of interest. A clear advantage of this strategy over existing methods that require cell isolation is that it does not involve tissue or animal dissociation into single cells. Tissue dissociation methods often require a long time to find an optimal condition to isolate cells of interest, maintaining cell viability. Even if such a condition is found, there is no guarantee that the isolated cells still retain the native transcriptome before the cell isolation step. Also, cell isolation methods are generally time-intensive and require expensive equipment to perform.

There are other cell-isolation-independent, yet pull-down-dependent methods to study RNA in a cell-specific manner. Transgenic expression of a tagged RNA-binding protein in specific cells followed by co-purification of the protein and RNA bound to it, enables to isolate RNAs involved in a particular biological process in a specific cell type. Methods that employ tagged ribosomal protein (Hupe et al., 2014) and polyA-binding protein (PABP) (Hwang et al., 2017) have been established, and each method can be used to study mRNA being translated and alternative polyadenylation events, respectively. As these methods neither directly address the transcriptional rate nor capture the transcriptome, these methods and SLAM-ITseq are complementary to each other and can study cell-specific RNA dynamics from different perspectives.

Around the same time SLAMseq was developed, two alternative methods for identifying 4-thiouridine incorporations were reported: thiouridine-to-cytidine sequencing (TUC-seq) (Riml et al., 2017) and TimeLapse-seq (Schofield et al., 2018). Unlike SLAMseq, which introduces T>C base conversions in cDNA at the reverse transcription step, these two methods directly introduce uracil-to-cytosine (U>C) conversions at the positions of 4-thiouracil in RNA using osmium tetroxide (OsO_4) and ammonium in TUC-seq, and

2,2,2-trifluoroethylamine (TFEA) and sodium periodate ($NaIO_4$) in TimeLapse-seq. Direct introduction of base conversions on RNA could potentially be beneficial when using an RNA-seq method that is independent of reverse transcription (e.g. Nanopore direct RNA-seq). The conversion efficiencies of SLAMseq and TUC-seq were both confirmed to be equally high (>90% conversion rate), while TimeLapse-seq has about an 80% conversion rate based on a restriction digestion assay. A thorough comparison with the same starting material is needed to conclude which method is the most sensitive and specific method to detect 4-thio-UTP incorporations.

3.6.2 Limitations of SLAM-ITseq

Sensitivity

Unlike the conventional RNA-seq methods, which identify differential gene expression based on the number of reads obtained for each gene, SLAM-ITseq needs to detect the difference in base conversion rates between samples. Although, as shown in the experiments using *Tie2-Cre*, SLAM-ITseq was confirmed to label housekeeping genes in endothelial cells that account only for <5% of the tissue, it is unknown if this method is sensitive enough to detect the difference in labelling levels of transcripts in even more minor types of cells in a tissue.

Also, metabolic RNA labelling level in a given exposure time is critically affected by two factors: (i) intracellular concentration of the analogue and (ii) transcriptional rate of each transcript. Hence, SLAM-ITseq might not be able to sensitively capture transcripts in cells that have less access to blood (e.g. neurons) or transcripts that have extremely long turnover times (e.g. rRNA).

A potential solution to capture transcripts with a low level of 4-thio-UTP incorporation would be to perform an enrichment step of the labelled transcripts prior to SLAMseq. Thiol-specific biotinylation followed by a streptavidin pull-down would allow us to perform deeper sequencing of the labelled transcripts without wasting sequencing power on reads from the majority of unlabelled transcripts. Also, its combination with SLAMseq would identify truly labelled transcripts from the unlabelled RNAs that are pulled-down non-specifically. Although the combination of streptavidin pull-down and SLAMseq has not yet been tested, TimeLapse-seq has employed this strategy and obtained more reads from nascent transcripts (e.g. pre-mRNA) (Schofield et al., 2018). However, this strategy has not yet been tested for its recovery rate (i.e. how much labelled transcripts are retained after the purification), and thus needs an extensive comparison to conclude if this is an appropriate method to increase the sensitivity.

Another limitation in sensitivity with SLAMseq is that its reliance on T>C conversions to discover labelled transcripts. As summarised in Chapter 1, RNA-seq inherently introduces errors in base calling at an order of 0.1% frequency. Thus, in theory, if the T>C conversions introduced by the incorporation of U analogues occur less frequently, this signal could be masked by the background error induced by the sequencing error. A simple solution to this issue is to use a long read paired-end sequencing. Since the same base is read twice, the probability of having the same T>C mismatch from the sequencing error should be less than once in 10^6 bases. This approach, however, will not account for the errors introduced during library preparation. Another solution would be to introduce unique molecular identifiers (UMIs) at the end of reads. Although this approach requires to redesign the library preparation method, it will find T>C mismatches introduced during PCR amplification by finding a consensus sequence among the reads with the same UMI.

Effects on RNA metabolism

Since our knowledge about effects of RNA modifications on RNA metabolism is still limited, it is still unknown if the 4-thiouracil-containing RNA retains the same biological properties in comparison to the native RNA. A study addressed this issue by culturing cells with the presence of different concentrations of 4-thiouridine and assessed cellular phenotypes and transcriptional changes (Burger et al., 2013). This study discovered that high concentrations of 4-thiouridine inhibit the synthesis of rRNA and decreases cell viability. Although, the *in vivo* concentration of 4-thiouracil we use in SLAM-ITseq is much lower than that was used in this *in vitro* experiment, the biochemical properties of RNA obtained from metabolic labelling experiments should be viewed with care.

Cre promoter specificity

As the cell-type-specificity of this method is dependent on the specificity of a promoter that drives Cre expression, labelling specificity with SLAM-ITseq is dependent on the Cre promoter expression specificity. This limitation applies to all methods that employ a Cre promoter to achieve cell-specificity and is not limited to SLAM-ITseq. Since some promoters are less cell-specific and sometimes even stochastic (Heffner et al., 2012), it is critical to choose a well-studied and reliable Cre promoter to achieve specific RNA labelling when performing SLAM-ITseq.

Also, if cells of interest cannot be selected with a single promoter (e.g. immune cells), FACS could still be an advantageous option as it can sort the cells with multiple markers.

Chapter 4

Analysis of mobile RNA in *M. musculus* with SLAM-ITseq

4.1 Background

Since the discovery of RNA as an essential biological molecule in the cell, studies have been conducted focusing on cell-autonomous RNA functions. However, in the past few decades, some compelling lines of evidence suggesting the mobility, and non-cell-autonomous roles of RNA in animals and plants have been shown. In this chapter, previous publications suggestive of RNA mobility are summarised, and the mobility of RNA in mouse is assessed with SLAM-ITseq.

4.1.1 Mobile RNA in nematodes

One of the most compelling lines of evidence that support the mobility of RNA was presented in 2001. A newly established model organism at that time, *Caenorhabditis elegans* (*C. elegans*) (Brenner, 1974), was shown to be effective in studying gene functions through transcriptional perturbation by short RNA targeting it, which is called RNA interference (RNAi) (Fire et al., 1991, 1998). Fire and colleagues showed that this RNAi can be induced not only through an injection of RNA but also through feeding bacteria that produce double-stranded RNA (dsRNA) (Timmons et al., 2001). Surprisingly, although dsRNA is taken up by intestinal cells, RNAi is induced systemically, and also in subsequent generations through germline. This result strongly suggests the spread of RNAi signal from intestinal cells to other cells including germ cells.

Genes involved in this systemic RNAi process have been identified through forward genetic screens. *C. elegans* expressing GFP only in body wall muscles and dsRNA targeting

GFP only in pharynx was generated, and the systemic RNA interference deficient (*sid*) phenotype was assessed. Forward genetic screens have identified SID genes involved in the *sid* phenotype (Winston et al., 2002). Among them, SID-1 and SID-2 have been well described. A transmembrane protein SID-1 has been suggested to be a dsRNA transporter across the cell membrane, and overexpression of SID-1 in *Drosophila* S2 cells increased the efficiency of dsRNA uptake by the cells (Feinberg and Hunter, 2003). SID-2, on the other hand, is highly expressed in intestinal epithelium and involved in environmental RNA uptake (Winston et al., 2007). Unlike SID-1, however, SID-2 is not involved in the spread of RNAi among different cells.

Although SID-1 is conserved across different animal kingdoms including mammals, it is important to note that the existence of SID-1 or SID-2 homologues in the genome is not sufficient for RNA mobility. For example, *Caenorhabditis briggsae* (*C. briggsae*) possesses both SID-1 and SID-2 in the genome, but is not able to induce systemic RNAi (Winston et al., 2007). Similarly, other *Caenorhabditis* nematodes have been tested for their ability of systemic RNAi through ingestion and injection (Nuez and Félix, 2012). The study revealed that both the uptake and spread of RNAi signal in *Caenorhabditis* genus are under rapid evolution, and only a subset of them shows such phenomena despite the high conservation of SID genes.

4.1.2 Mobile RNA in plants

An elegant experiment in plants also showed the mobility of RNA between different parts of a plant. It has been known that the transgenic tabacco line homozygous for 35S-*Nia2* transgene exhibits silencing of both the host and transgenic *Nia2* (Palauqui et al., 1996). Also, this co-suppression of the genes is observed in a constant fraction of homozygous progeny. Thus, using this system, isogenic plants with or without the silencing effect can be obtained, and each state is called as S (silenced) and NS (non-silenced), respectively. By grafting NS scion on S stocks, silencing of *Nia2* in the grafted NS scion was observed, suggesting that the silencing signal spreads from the stock to the scion (Palauqui et al., 1997) (Fig. 4.1). Also, this spread of the silencing signal was discovered to be phloem-dependent.

The short-distance spread of transgene silencing signal has also been discovered in *Nicotiana benthamiana* (Voinnet and Baulcombe, 1997). By delivering a full-length GFP transgene by *Agrobacterium tumefaciens* to a part of a plant bearing a promoter-less GFP transgene, gradual silencing in the inoculated part was induced, suggesting that the GFP silencing signal spread from neighbouring parts of the plant (Voinnet et al., 1998).

Both long- and short-range spread of transgene silencing revealed to involve small RNA species (Dunoyer et al., 2007). Northern blot analyses discovered two classes of

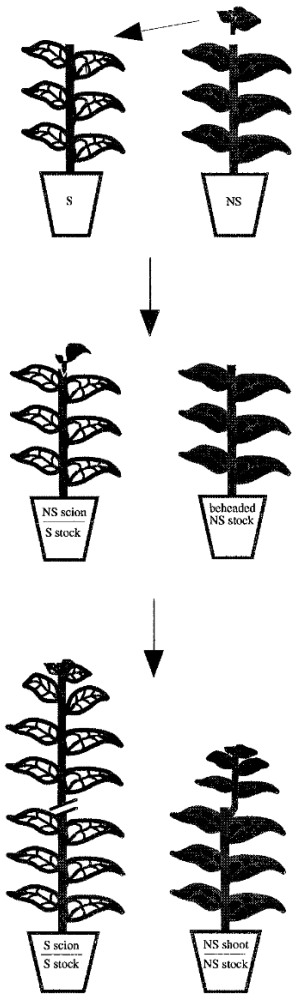


Fig. 4.1 The graft experiment that showed the spread of gene silencing signal in plants Schematic of the design of the experiment that showed the intercellular silencing signal in plants (taken from Palauqui et al. (1997)). Non-silenced (NS) scion was grafted on silenced (S) stocks, and the gene expression in the graft and the host was assayed. Both plants were kept and observed to carefully exclude the possibility that the transgene silencing was induced by time.

siRNA, 21-22 nt and 25 nt long, that silence a transgene. The mobile siRNA research in plants has experienced a shocking news that reported that two groups have admitted image manipulations, following retractions of several key publications in the field. However, the mobility of siRNA has been confirmed by independent groups (Molnar et al., 2010), suggesting that the scientific claims made in the retracted publications may still be valid.

Several miRNA species were also suggested to be mobile between different parts of plants similar to siRNA. Still, for both siRNA and miRNA, the detailed mechanisms of action need to be explored. Also, more importantly, functional consequences of the small RNA mobility is still elusive.

4.1.3 Extracellular RNA in mammals

In mammals, the studies on RNA mobility began with the discovery of extracellular RNA. Scientists started to analyse extracellular nucleic acid concentration as early as 1931 (Javillier and Fabrykant, 1931), when functions of DNA and RNA were not yet known. These studies were conducted in the context of studies on how different metabolites are exchanged between cells and plasma. Concentration of DNA and RNA in human plasma was determined with rapid inhibition of ribonuclease (RNase) (Kamm and Smith, 1972) in order to prevent degradation of the nucleic acids.

Extracellular RNA research has became a hot topic when tumour-derived RNA was detected in human plasma (Kopreski et al., 1999; Lo et al., 1999), as it promised discovery of novel diagnostic markers. A characterisation of circulating RNA in the plasma revealed that the majority of the serum RNA is of low molecular weight. Deep sequencing of small RNA in plasma revealed that abundant miRNA is stably present in the plasma (Mitchell et al., 2008).

The big remaining question is how the extracellular RNA is protected from degradation. Since RNase exists in extracellular space, any naked RNA in the circulation will be degraded rapidly. To test if the circulating RNA is associated with any particles that protect RNA from degradation, RNA concentration was measured after filtering with different sizes of pores (Ng et al., 2002). This experiment indeed suggested the possible existence of RNA that is associated with particles. Also, though extracellular RNA is stable for at least a few hours in plasma at room temperature, an addition of detergent to plasma induced a rapid degradation of the RNA, suggesting that extracellular RNA is protected by either lipid bilayer or protein (El-Hefnawy et al., 2004).

An experiment identified that small (50-90 nm) extracellular vesicles called exosomes are present in the serum and small RNAs were discovered in these vesicles. Since then, the vesicle has become a strong candidate that keeps RNA intact in the extracellular space (Valadi

et al., 2007). On the other hand, RNA-binding proteins have also been proposed to protect RNA from degradation (Arroyo et al., 2011). In this report, an assay was performed to identify the protective factor of RNA by exposing extracellular fluid to different treatments, detergent or protease, and RNA abundance in the fluid after the treatment was compared. While the protease treatment significantly reduced the concentration of extracellular RNA, the detergent treatment did not, which suggests that the majority of RNA is bound by RNA-binding proteins and thereby remains stable in the extracellular space. A miRNA-binding protein, Ago2 pull-down experiment in plasma identified abundant miRNAs bound to Ago2. Yet, since no biological functions have been discovered for extracellular RNAs, there is no consensus on which mode of protection is more biologically relevant.

4.1.4 Mobile RNA in mammals

As the research on extracellular RNA grew, some started to hypothesise that these extracellular RNAs are mobile between different cells and function as an intercellular signal. A number of experiments were performed to test this hypothesis by transferring purified exosomes from a dish of cultured cells to another, and mRNAs derived from the donor cells were observed in the recipient cells (Valadi et al., 2007). Although some of these experiments are suggestive of RNA transfer through exosomes, it is hard to conclude that the same is true *in vivo*, as these experiments were mostly performed using immortalised cells, and the exosome concentration introduced in the medium was presumably much higher than the physiological concentration.

An elegant approach was invented to study the mobility of transgene-derived mRNA *in vivo* by using mice with a Cre-inducible LacZ transgene and a Cre transgene under a cell-type-specific promoter. With this approach, the mobility of Cre mRNA can be tested: if there is mobile Cre mRNA to a non-Cre-expressing cell, the recipient cell will be detectable through the LacZ positive signal. Ridder et al. generated mice expressing Cre in the haematopoietic lineage, and, surprisingly, detected LacZ-positive neurons in the brain. They carefully excluded the possibility that this was caused due to cell fusions through immunofluorescence assays. Hence, this suggests the mobility of Cre mRNA from haematopoietic cells to neurons. However, since this experiment does not show the mobility of endogenous transcripts, and the number of LacZ-positive neurons detected was quite low, it is hard to link this phenomenon to any physiological processes. Also, it could just be an ectopic expression of the Cre in the neurons.

Below, some experiments that attempted to show the mobility of endogenous RNA are summarised.

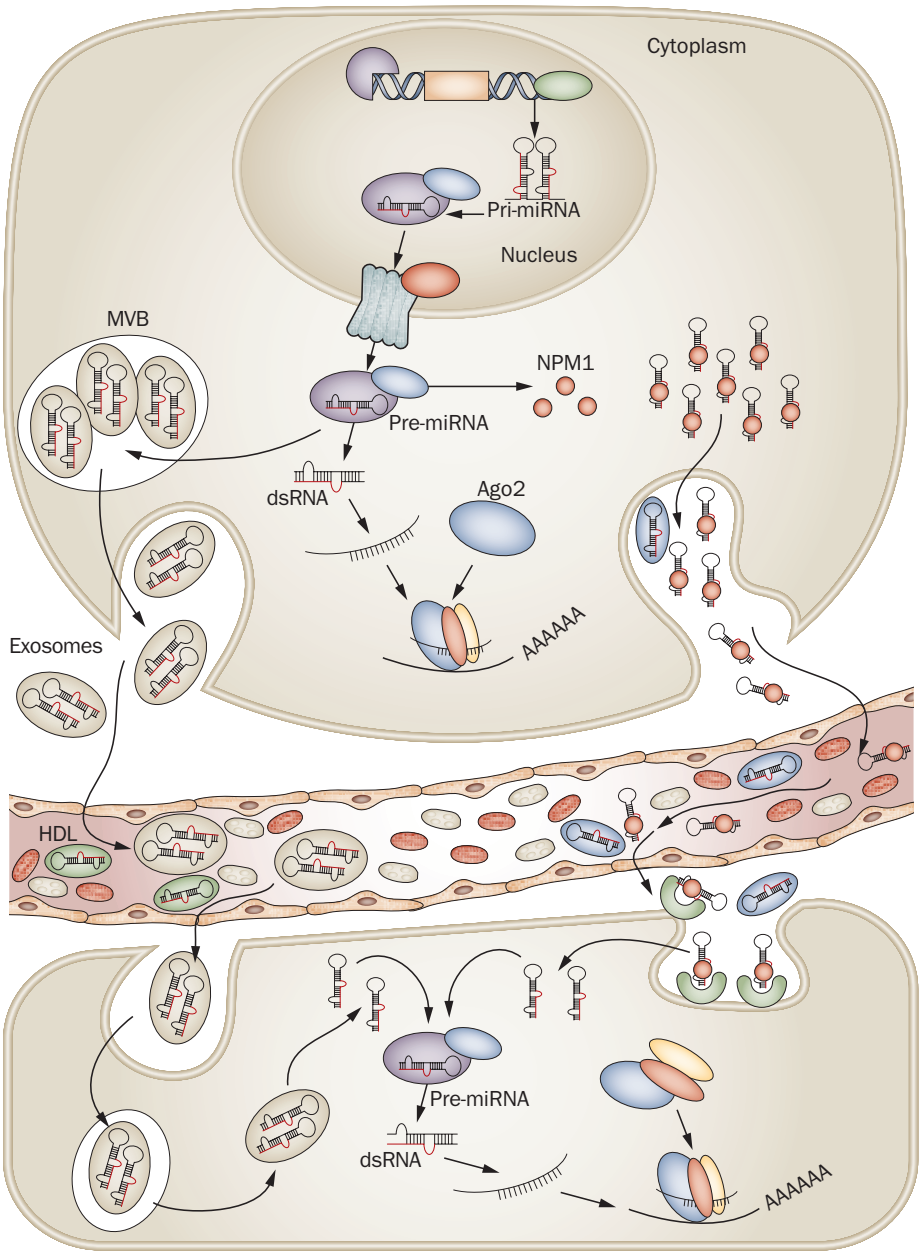


Fig. 4.2 Working hypotheses of miRNA mobility in mammals
Schematic summary of potential pathways, in which miRNAs are transferred from one cell to another. Pre-miRNAs synthesised in a donor cell are released into the circulation and are transferred to a recipient cell contained in exosomes or by proteins bound to them. The pre-miRNAs delivered to the recipient cell are further processed by Dicer to generate mature miRNAs and regulate gene expression in the recipient cell. Taken from Maria Angelica et al. (2011).

Adipose-to-liver RNA transfer

A series of experiments have been performed to test if adipose-derived miRNAs are released into the circulation and if they are delivered to liver (Thomou et al., 2017). First, adipose-specific Dicer knock-out mice (ADicerKO) were generated, and the abundance of circulating miRNA was compared with wild-type mice. The overall abundance of circulating miRNA was significantly lower in the ADicerKO, which suggests that adipocytes are the major contributer to the circulating miRNA pool. Next, the study assessed the intercellular mobility of miRNAs *in vivo*. They injected adenovirus bearing a human pre-miR302f directly into brown adipose tissue (BAT), and, four days later, another adenovirus bearing a luciferase reporter attached to the $3'$ UTR sequence that miR302f binds to was introduced intraveneously so that it was delivered to liver. The luciferase reporter expression in the liver was lower in the pre-miRNA-injected mice compared to control, suggesting that the miR302f was transferred from BAT to liver and suppressed the expression of the reporter. Although this direct assessment of mobility *in vivo* paves the way to understanding miRNA mobility in mice, this particular experimental result could be just because the adenovirus bearing pre-miRNA was delivered to liver. Also, even if this particular miRNA is mobile as stated, this study raised new questions: how general is this phenomenon among different species of RNA and cell types? What is the physiological significance of this RNA mobility?

Gut-to-liver RNA transfer

A potential communication between intestine and liver was also shown (Deng et al., 2013). Intestinal epithelial cells are known to secrete extracellular vesicles. Deng et al. collected the extracellular vesicles from intestine, labelled them with infrared dye, and administered the labelled vesicles to another mice orally. They showed that fluorescent signal was detected in the intestine and liver of the recipient mice, which shows the mobility of the administered vesicle from intestine to liver. Although this experiment proves neither the mobility of vesicles synthesised in intestinal cells nor the uptake of the contents of the vesicles in liver, the proposed hypothesis seems reasonable considering the fact that blood flows from intestine to liver through the portal vein.

Epididymis-to-sperm RNA transfer

The majority of steps of spermatogenesis takes place in the testicular tubule, and when the cells become elongating spermatids, they are released to another tubular organ called epididymis attached to testis. The epididymis was once thought to just store sperm before ejaculations, but it turned out to be important for sperm to get fully matured. The epididymis

can anatomically be divided into three parts from the proximal part: caput, corpus, and cauda epididymis. Sperm collected from the caput shows lower mobility and less efficient acrosome reaction with oocytes compared to the ones collected from the cauda (Lakoski et al., 1988). Studies have found that the epididymal epithelium releases extracellular vesicles named epididymosomes, which contain various biological molecules that are thought to be important for sperm maturation (Frenette et al., 2002). Co-incubation of sperm and epididymosomes revealed that they can fuse together.

Recent studies have found that epididymosomes also contain small RNA species, and two papers claim that RNA is transferred from the epididymis to sperm using epididymosome as a cargo (Chen et al., 2016; Sharma et al., 2016). More surprisingly, the transferred RNA was suggested to regulate the transcriptional network in the mouse embryo and, in turn, affects phenotypes of the offspring. Mechanisms of the transfer and mode of actions of the small RNA in embryos still need further investigations.

Also, the same group reported that this RNA transfer is even essential for mouse development (Conine et al., 2018). They showed that the sperm collected from the caput epididymis is not only incapable of fertilisation but also incapable of progressing development. The lower developmental potential observed from the caput sperm-derived embryos was rescued by exposing the sperm to epididymosomes prior to fertilisation, suggesting that the molecules contained in epididymosomes are critical for the embryonic development. However, recently, another group sent a letter to the journal claiming that they successfully generated mice from caput sperm, and thus epididymosomal transfer is not essential for mouse development (Zhou et al., 2019). Also, in human, *in vitro* fertilisation (IVF) is routinely performed using sperm obtained from epididymis, and even the testicular sperm is also capable of progressing development. Thus, it does not seem to be reasonable that murine sperm loses the developmental potency only while they are in the caput epididymis.

4.1.5 Mammalian SID-1 orthologs

Since SID-1 is suggested to be an RNA transporter in *C. elegans*, a number of studies have analysed the functions of its orthologs, called Sidt1 and Sidt2. Based on the RNA expression atlas[1], both genes show a general expression pattern across mammalian tissues, but *Sidt2* has higher expression compared to *Sidt1* in many tissues.

A number of *in vitro* studies have addressed the functions of Sidt1. One study has reported that the overexpression of Sidt1 in mammalian cells enhances the cellular ability to uptake RNA and thus is useful for RNAi experiments (Wolfrum et al., 2007). The functional

[1] https://www.proteinatlas.org

significance of Sidt1 *in vivo* has also been assessed through generating KO mice. The KO mouse database shows that Sidt1 KO mice had been generated and their phenotypes were assessed, but no clear phenotypes were observed[2]. Independent of this large-scale KO project, a group also generated Sidt1 KO mice and assessed the ability of RNA transfer in immune cells (Nguyen et al., 2019). They identified that Sidt1 localises to the endosomal and lysosomal membrane and is involved in releasing dsRNA to the cytosol through direct interaction with the RNA.

Involvement of Sidt2 in RNA uptake has also been studied, and a group reported that the overexpression and knock down of Sidt2 led to increased and decreased uptake of single-stranded oligonucleotides, respectively (Takahashi et al., 2017). Sidt2 KO mice have also been generated, and several KO-associated phenotypes have been reported. Metabolic alterations including glucose intolerance and changes in liver histology were observed in the KO mice, suggesting the link between Sidt2 and liver functions (Chen et al., 2018; Gao et al., 2013, 2016), though the molecular mechanisms that cause these phenotypes are unclear. Also, another group assessed the significance of Sidt2 in immune cells, where relatively higher expression of Sidt2 is observed (Nguyen et al., 2017). They showed that the cellular ability of dsRNA uptake was unchanged between KO and wild-type (WT), and that dsRNA was mostly confined in the endosomes in the KO mice, whereas it was diffused in the cytosol in WT. These observations suggest that Sidt2 is not involved in the dsRNA uptake but is important for dsRNA release from the endosomes. Further, the KO mice were fed with human sarcoma virus (HSV), which is an RNA virus, and lower survival was observed in the KO mice compared with WT. Thus, these results show for the first time that transmembrane RNA mobility in immune cells are important for antiviral immunity.

However, there is also a contradictory report about human SIDT1 and SIDT2. Valdes et al. discovered a gene involved in cholesterol transport in *C. elegans* and named it CUP-1. Surprisingly, their *in silico* sequence analysis revealed that human SIDT1 and SIDT2 are closer homologues to CUP-1 than SID-1. Also, they found that the expression of neither SIDT1 nor SIDT2 is linked with the cellular ability of dsRNA uptake *in vitro*, but, instead, that SIDT1 overexpression increased cholesterol uptake. Interestingly, when dsRNA was provided with cholesterol, the overexpression of SIDT1 increased dsRNA uptake. Since the previous papers reporting the significance of SIDT1 and SIDT2 in RNA uptake were shown using lipophilic RNAs, the authors speculate that mammalian SIDTs are just involved in cholesterol transport but not RNA transport. Also, this hypothesis fits well with the metabolic phenotypes observed in the Sidt2 KO mice, though it does not explain the RNA release from endosomes observed in the immune cells.

[2]https://www.mousephenotype.org/data/genes/MGI:2443155

4.1.6 Mobile RNA and intergenerational epigenetic inheritance

The mobile RNA phenomenon is particularly fascinating when it is viewed in the context of information transmission from soma to germline.

Whether parental experiences affect offspring has been a great question since the 19th century. The theory about inheritance of acquired phenotypes proposed by Lamarck (Lamarck, 1809) has become unrecognised after being backlashed by Weismann both theoretically and experimentally (Weismann, 1892). Recently, however, some neatly designed experiments have shown that some phenotypic differences were observed when parents are exposed to certain stimuli (Ng et al., 2010). To best focus on the inheritance through germline, a number of experiments have been designed to expose male animals to stimuli, and the phenotypic changes in the offspring were assessed. Genome-wide epigenetic analyses on spermatozoa revealed that a wide range of molecular states in the spermatozoa was altered in response to the stimuli, which includes changes in DNA methylation (Carone et al., 2010; Radford et al., 2014) and abundance of RNA species (Gapp et al., 2014). Some research groups reported that differential RNA expression in sperm is sufficient to induce the phenotypic changes observed by showing the reproduction of the phenotypes through injecting RNA obtained from the sperm of males that were exposed to the stimuli to zygotes (Chen et al., 2016; Gapp et al., 2014; Sharma et al., 2016).

A common working hypothesis in the field is that somatic cells exposed to a certain stimulant release RNA that is later taken up by germ cells, which in turn alters the phenotypes of offspring derived from the germ cells. This hypothesis is surprisingly similar to the conceptual substance called “gemmules” proposed by Charles Darwin (Darwin, 2010), although, unlike the hypothetical gemmules, extracellular RNA or exosomes will not form germ cells themself.

4.2 Aims of this chapter

Although a number of studies have been conducted to assess the mobility of RNA between mammalian cells, it has been challenging to directly assess the mobility of RNA *in vivo*. Most of the studies have utilised *in vitro* systems, in which medium or extracellular vesicles collected from cells were transferred to another. Although a few *in vivo* experiments have been performed, they mostly showed the mobility of transgene-derived RNA, and thus it is still unclear if there is any endogenous mobile RNA in mammals.

To directly and comprehensively assess the intercellular mobility of endogenous RNA in mice, I employed SLAM-ITseq. By generating mice expressing UPRT in a specific cell type, potential “donor” cells, I tested if the labelled RNA is mobile to any other tissues by detecting

the labelled RNA in non-UPRT-expressing cells, potential "recipient cells" (Fig. 4.3). Tissues for the analyses were chosen based on previous publications suggesting mobile RNA and the availability of well-studied Cre lines.

To optimise the RNA labelling time *in vivo*, 4-thiouridine exposure to WT mice was first performed instead of the transgenic mice. After determining the exposure time that is sufficient to label small RNAs, the same exposure time was used for mobile RNA assays with cell-type-specific RNA labelling animals.

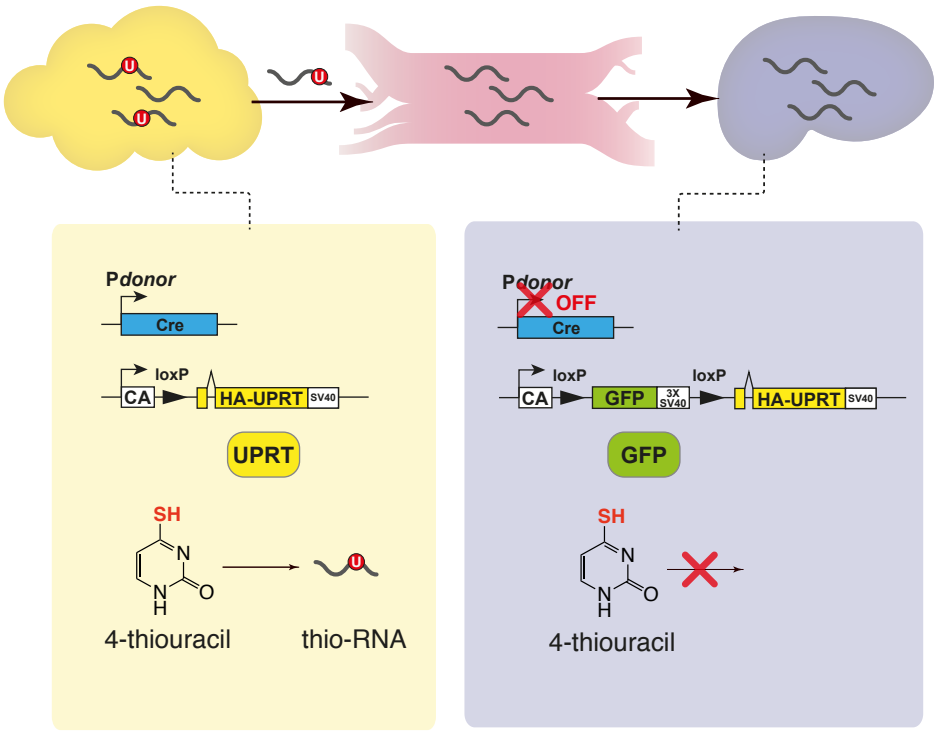


Fig. 4.3 Schematic representation of the experimental design to detect mobile RNA *in vivo*

Transgenic mice expressing UPRT in a specific cell type (i.e. donor cells) is generated and exposed to 4-thiouracil to label RNA synthesised in the cells. RNA collected from the serum and non-UPRT-expressing tissues (i.e. recipient tissue) was analysed with SLAM-ITseq to see if the RNA labelled in the donor cells are detectable. Wavy line, RNA; red circle, incorporated 4-thiouracil; P*donor*, a promoter specific to the donor cell.

4.3 Confirmation of extracellular small RNA labelling with 4-thiouridine

In order to achieve *in vivo* metabolic labelling to detect mobile RNA in mice, I first performed 4-thiouridine injections to WT mice to find an optimal dosing condition for the experiment. Since 4-thiouridine is incorporated into RNA independent of UPRT, all the cells exposed to 4-thiouridine should synthesise the labelled RNA. The thiol-labelling level of the circulating miRNA is used as a readout to assess if a given exposure method to an analogue is sufficient to observe mobile RNA. As 4-thiouracil and 4-thiouridine can both be administered in the same way, they presumably have similar physiological dynamics *in vivo*.

Considering many miRNA species have a half-life of 24 h or longer (Duffy et al., 2015), an exposure time of 54 h was used to maximise the labelling level. Three injections were performed with 24 h intervals, and tissues were collected 6 h after the last injection (Fig. 4.4). This shorter interval of 6 h between the last injection and tissue collection was chosen in order to capture miRNAs with a shorter half-life before they were degraded. WT mice were exposed to either 4-thiouridine or DMSO control, and RNA was extracted from the serum. To capture both exosomal and non-exosomal miRNAs, RNA was isolated from the serum without any fractionation. The alkylation reaction with IAA was performed as described before, and SLAMseq was performed on the IAA-treated small RNA.

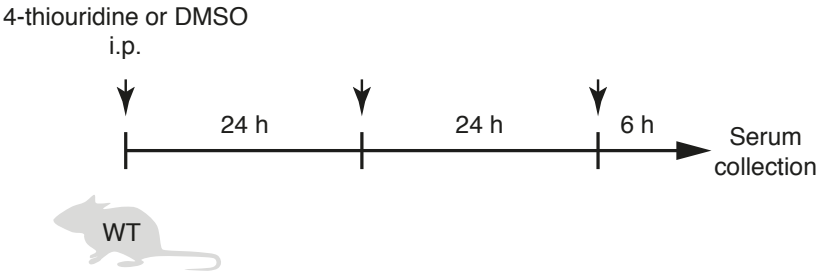


Fig. 4.4 4-thiouridine injection scheme to label circulating miRNA
WT mice were i.p. injected with 4-thiouridine or DMSO (N = 3 each). Two additional injections were performed at 24 h intervals, and the serum was collected 6 h after the last injection.

To identify circulating miRNAs that were labelled by the 4-thiouridine injections, miRNA species with higher T>C conversions in 4-thiouridine serum were determined by comparing the T>C conversion rates between 4-thiouridine and DMSO mice (Fig. 4.5A). The miRNA expression level was also compared between 4-thiouridine and DMSO mice, and a high positive correlation of the abundance of miRNAs was confirmed, suggesting that 4-thiouridine incorporation did not have a significant impact on serum miRNA abundance (Fig. 4.5B).

Although a higher proportion of highly expressed miRNAs were labelled, lowly expressed miRNAs were also labelled (Fig. 4.5C), suggesting that wide-ranged circulating miRNAs were labelled.

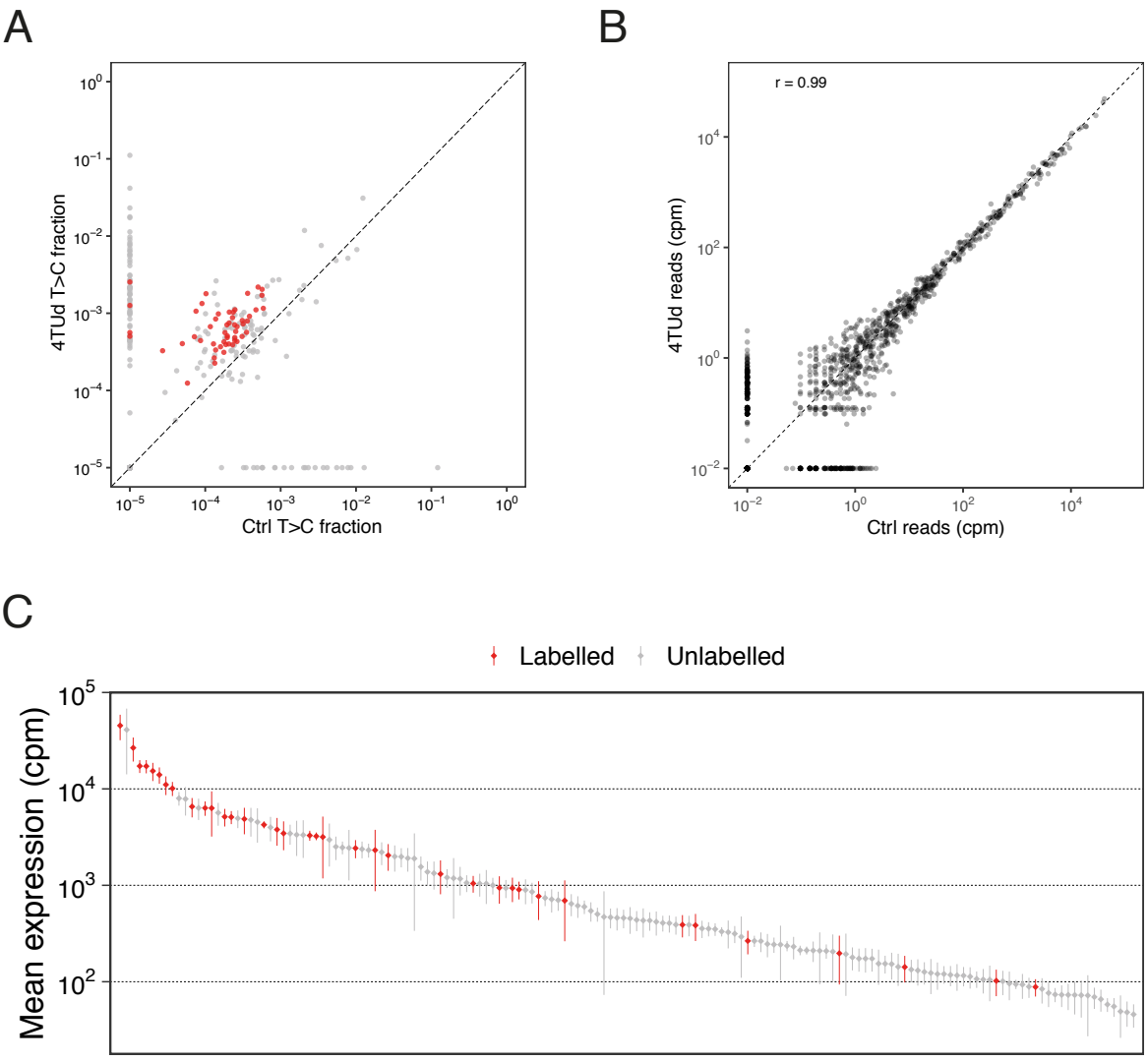


Fig. 4.5 Circulating miRNAs labelled by 4-thiouridine injections
(A) T>C conversion rate of the serum miRNAs comparing 4-thiouridine and DMSO mice. miRNAs with significantly higher T>C rate in the 4-thiouridine-exposed serum are shown in red (beta-binomial test FDR <0.1; N = 3 for each condition). A constant value of 10^{-5} was added when plotting. (B) Serum miRNA expression in 4-thiouridine and DMSO mice is compared. Spearman's correlation coefficient between the two conditions is shown. A constant value of 10^{-2} was added when plotting. (C) miRNAs that were detected in all samples are sorted by their mean expression among all samples (N = 6). Significantly labelled miRNAs are shown in red. Each point represents the mean expression of miRNA among all samples, and the vertical lines show 95% confidence intervals.

Since 4-thiouridine can only be incorporated into genomic T positions of miRNAs, the T content of labelled and total miRNA was compared to test if miRNAs with high T content were enriched in the labelled fraction (Fig. 4.6). However, distributions of the labelled and total miRNAs with different genomic T content do not differ significantly, which suggests that labelled circulating miRNAs identified with SLAMseq did not enrich for miRNAs with higher T content.

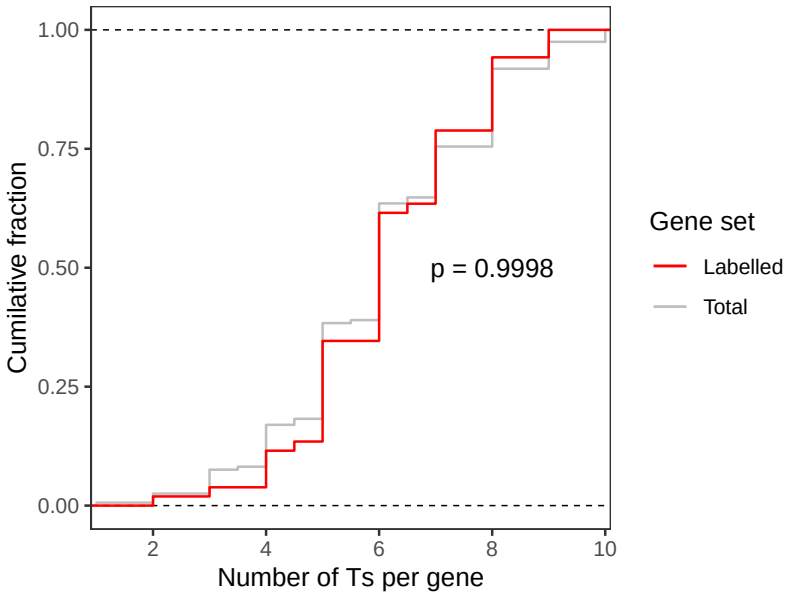


Fig. 4.6 Cumulative distribution of the number of Ts in labelled and total miRNAs
Empirical cumulative distribution function plot of the number of Ts in labelled and total circulating miRNAs. The *P*-value obtained from the two-tailed KS-test is shown.

Circulating miRNAs are thought to be originated from multiple tissues in a body. To confirm if the 4-thiouridine injections labelled circulating miRNAs released from specific tissues with bias, a list of the labelled miRNAs by 4-thiouridine is obtained, and their expression patterns in different tissues were assessed using a published miRNA expression atlas (de Rie et al., 2017). A heat map of the cellular expression pattern for the labelled miRNA shows that some labelled miRNAs have a broad expression pattern across cell types, while others have a highly specific expression pattern. However, no obvious cluster in any specific cell types was found, and thus it suggests that the systemic 4-thiouridine injection labelled the circulating miRNAs with little bias towards any miRNAs originated from specific tissues.

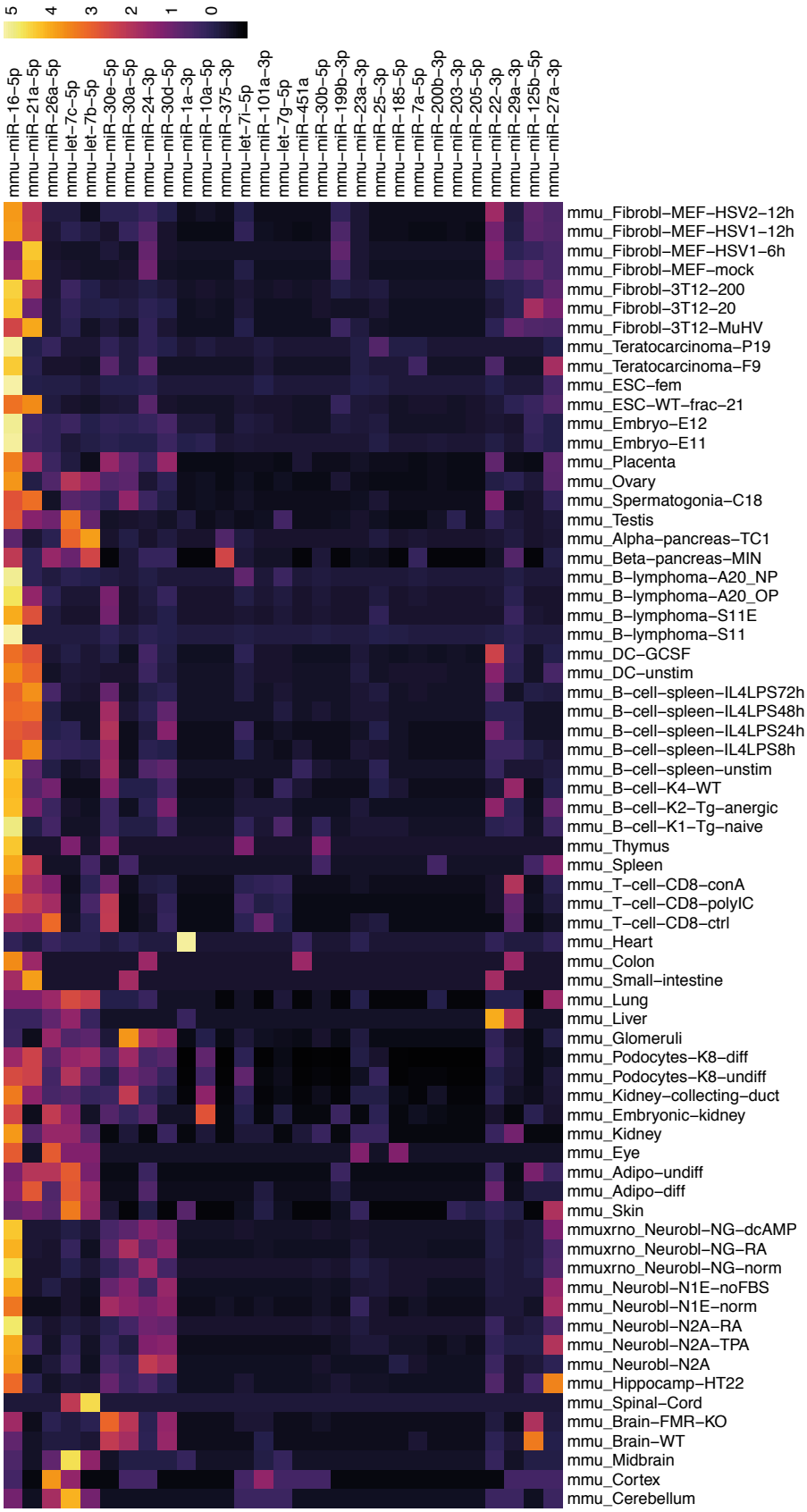


Fig. 4.7 Heat map showing expression patterns of the labelled miRNAs in different murine cell types
miRNA expression data was obtained from de Rie et al. and summarised as a heat map for the labelled miRNA identified. The colour key represents Z-score across the rows.

4.4 Mobile RNA assay with SLAM-ITseq

Now that this 3-time injection of 4-thiouridine was confirmed to label circulating miRNAs with little bias, the same protocol was applied to the SLAM-ITseq method to study the mobility of miRNA synthesised in particular cell types.

Both Cre^+ and Cre^- mice were exposed to 4-thiouracil by i.p. injections, and RNA from the tissue containing UPRT-expressing cells, serum, and a non-UPRT-expressing tissue was collected. SLAMseq was performed on the isolated RNA, and labelled miRNAs were determined based on the T>C rate difference between Cre^+ and Cre^-.

Based on the previous studies that showed the intercellular mobility of transgene-derived RNA or exosomes, three pairs of tissues were chosen for the analyses: WAT to liver (Thomou et al., 2017), intestine to liver (Deng et al., 2013), and epididymis to sperm (Chen et al., 2016; Sharma et al., 2016, 2018). To achieve this goal, three different Cre mice were used to generate both Cre^+ and Cre^- mice: *Adipoq-Cre* (adipocyte specific), *Vil-Cre* (gut epithelium specific), and *Spink8-Cre* (epididymal epithelium specific). The specificity of the expression of these promoters have previously been extensively tested, and *in vivo* labelling specificity using 4-thiouracil with *Adipoq-Cre*$^+$ and *Vil-Cre*$^+$ mice were confirmed in Chapter 3.

4.4.1 Adipose-to-liver RNA transfer was not detected

Adipoq-Cre was used to test if adipocyte-derived RNA is mobile. Epididymal white adipose tissue (eWAT) was chosen as a representative adipose tissue and was used to confirm if miRNA labelling in the donor cells was achieved. *Adipoq-Cre* should be expressed in both WAT and brown adipose tissue (BAT). SLAM-ITseq was performed on the small RNA extracted from eWAT, serum, and liver, and T>C rate for miRNAs was compared between *Adipoq-Cre*$^+$ and *Adipoq-Cre*$^-$.

Significantly labelled miRNAs were found in eWAT, which suggests that successful small RNA labelling was achieved in the donor tissue (Fig. 4.8A). To confirm the abundance of labelled and unlabelled miRNAs, the expression level of all the sequenced miRNAs are plotted (Fig. 4.8B). Strikingly, the majority of the highly abundant miRNAs were labelled, and some of the miRNAs with mid-low expression level were also labelled.

However, when the beta-binomial test was performed on the labelling level of miRNAs detected in the serum and liver, none of them reached FDR <0.1 (Fig. 4.8A). To compare the miRNA species sequenced from different tissues, an Euler digram summarising all the sequenced miRNAs in eWAT, serum, and liver was generated (Fig. 4.8C). Since there is a significant overlap of the detected miRNAs among these three tissues, non-detection of the

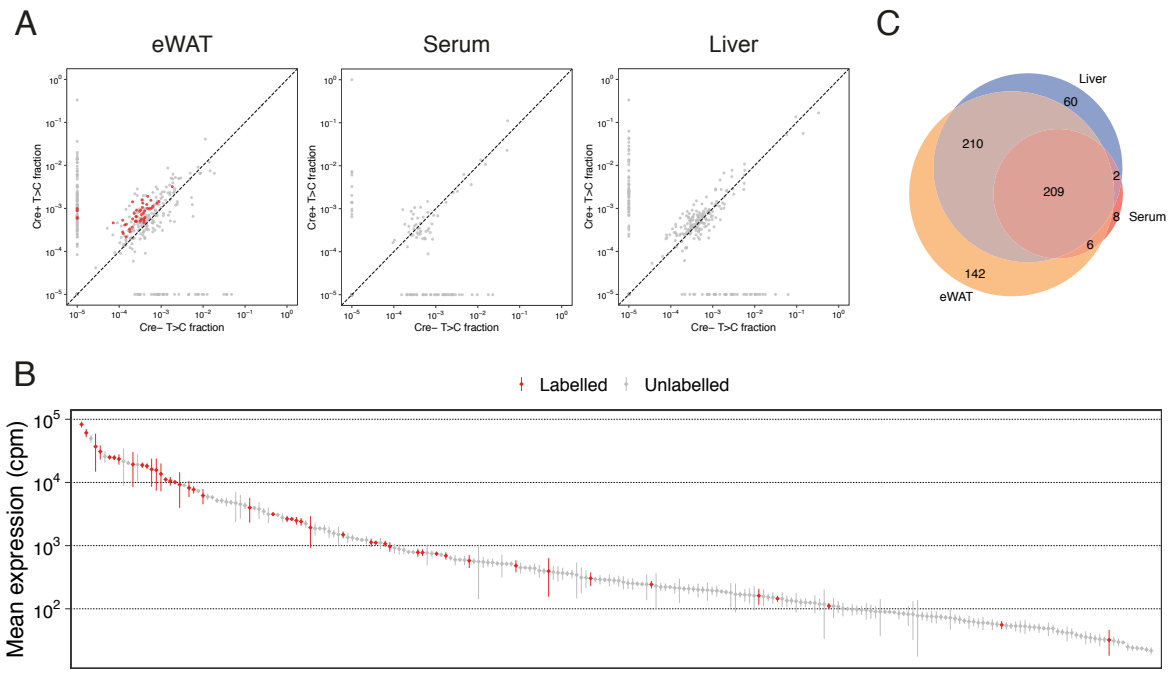


Fig. 4.8 Mobile RNA assay between eWAT and liver
(A) T>C rate of miRNAs between *Aipoq-Cre*$^{+}$ and *Adipoq-Cre*$^{-}$ was compared. Significantly labelled miRNAs are shown in red. A constant value of 10^{-5} was added when plotting. (B) The expression of the sequenced miRNAs is ordered by their mean expression and shown in descending order. Diamonds, mean expression among samples ($N = 6$); vertical lines, 95% confidence intervals; red, labelled miRNAs; grey, unlabelled miRNAs. (C) Euler diagram represents overlaps among the sequenced miRNAs in eWAT, serum, and liver.

labelled miRNAs in serum and liver is not because the labelled miRNAs were not sequenced in these tissues.

4.4.2 Intestine-to-liver RNA transfer was not detected

Exosomes taken up from intestine are known to be transferred to liver, and intestinal epithelium has the capacity to release extracellular vesicles containing RNA through *in vivo* and organoid studies (Deng et al., 2013; Szvicsek et al., 2019). To test if the RNA contained in these exosomes was released from the intestinal epithelium, SLAM-ITseq was used to label RNA in the intestinal epithelium and to test if the labelled RNA was found outside intestine. Villin is known to be expressed in all types of the intestinal epithelial cells (Madison et al., 2002), and thus *Vil-Cre*$^{+}$ mice were used to label RNAs in all the intestinal epithelium cells to see if any RNA generated in the cells is mobile to the serum or liver.

Similar to the experiment using *Adipoq-Cre*, *Vil-Cre*$^{+}$ and *Vil-Cre*$^{-}$ mice were exposed to 4-thiouracil for three times, and duodenum, serum, and liver were collected. RNA extracted from these tissues was used for SLAM-ITseq analyses to find labelled miRNAs. Significantly labelled miRNAs in *Vil-Cre*$^{+}$ mice were found by comparing the T>C ratio (Fig. 4.9). A plot summarising the expression level of miRNAs confirmed that most of the abundant miRNAs were significantly labelled (Fig. 4.9B).

However, similar to the previous experiment, no labelled miRNAs were found in the serum and liver. To confirm the similarity of the repertoires of miRNAs detected in these three tissues analysed, an Euler diagram comparing detected miRNAs in these tissues was generated (Fig. 4.9C). A considerable overlap is seen among the detected miRNAs in these tissues, suggesting that the labelled miRNAs in intestine were also sequenced in the serum and liver, and that non-detection of miRNAs in the serum and liver was not due to failure in RNA sequencing.

4.4.3 Epididymis-specific RNA labelling was achieved with *Spink8-Cre*

Although specific Spink8 expression in the epididymal epithelium has been reported (Jalkanen et al., 2006), the specificity of Cre expression in *Spink8-Cre*$^{+}$ mice has not been confirmed. To test the specificity of RNA labelling with this transgene, SLAM-ITseq was performed on the polyA RNA extracted from the cauda epididymis. The beta-binomial test identified genes with a significantly higher T>C rate in *Spink8*$^{+}$ mice (Fig. 4.10A). To confirm whether cell-type-specific RNA labelling was achieved, a few marker genes of epididymis as well as sperm were chosen as positive and negative controls, respectively (Fig. 4.10B). As expected, while known epididymal genes (*Spink8*, *Spink10*, and *Wfdc10*) were significantly labelled,

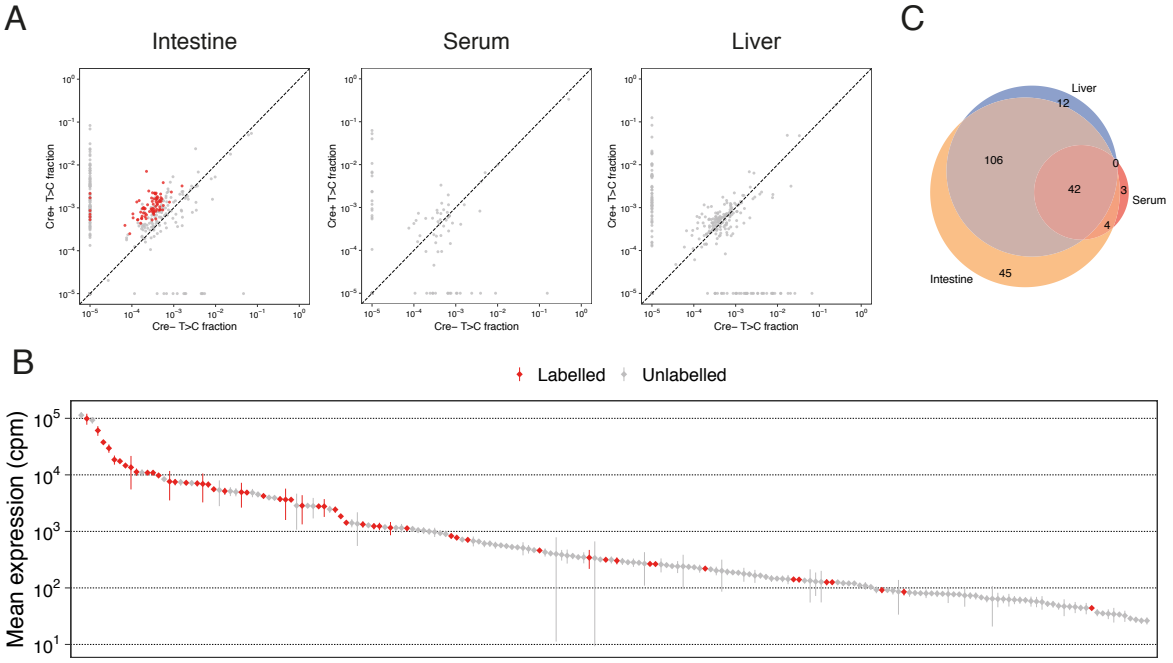


Fig. 4.9 RNA mobility assay between intestinal epithelium and liver
(A) T>C rate of miRNA in *Vil-Cre*$^{+}$and *Vil-Cre*$^{-}$ is compared in intestine, serum, and liver. Significantly labelled miRNAs are shown in red (beta-binomial test; FDR <0.1). A constant value of 10^{-5} was added when plotting. (B) Mean expression of the miRNAs detected in all the samples is sorted by the expression and is shown in descending order. Diamonds, mean expression of miRNAs among samples; vertical lines, 95% confidence intervals; red, labelled miRNAs; grey, unlabelled miRNAs. (C) Euler diagram representing overlaps of miRNAs detected in intestine, serum, and liver.

sperm genes (*Prm1*, *Prm2*, and *Tnp2*) were not labelled, suggesting that epididymal-specific RNA labelling was achieved.

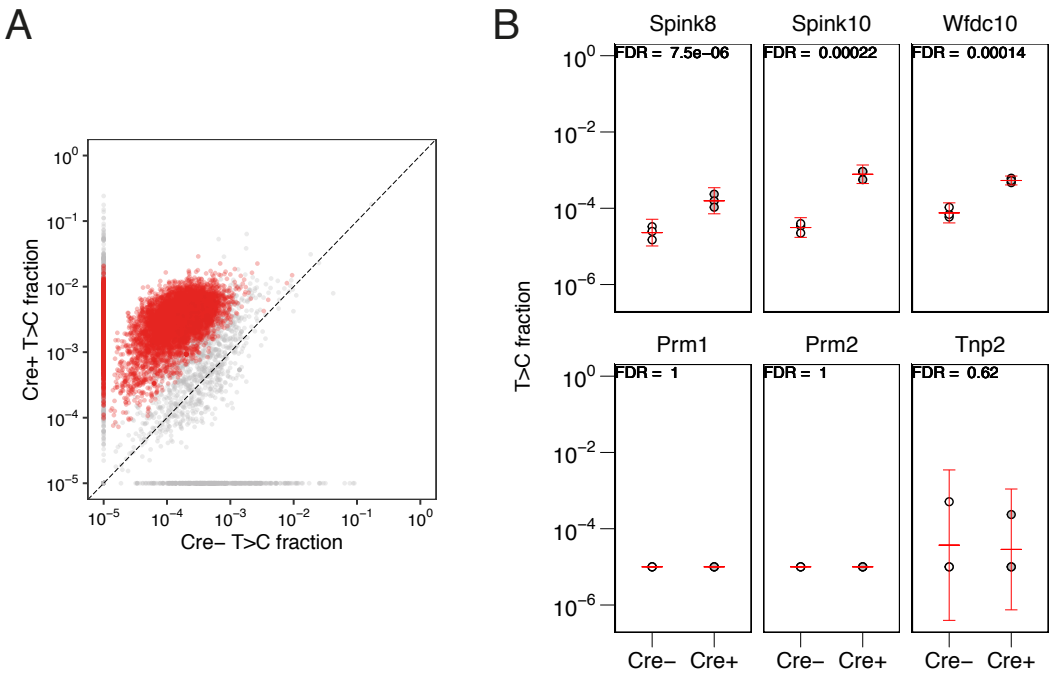


Fig. 4.10 Labelled transcripts in *Spink8-Cre*$^+$ were identified
(A) Mean T>C rate of each transcript in Cre$^+$ and Cre$^-$ is shown. Genes with a significantly higher T>C rate in Cre$^+$ are shown in red. A constant value of 10^{-5} was added when plotting. (B) T>C rate of each biological replicate for known epididymal genes (upper panes) and sperm genes (lower panes) are shown. A constant value of 10^{-5} was added when plotting. Red bars, mean and 95% confidence intervals across biological replicates (N = 3 each for Cre$^+$ and Cre$^-$); FDR was calculated by the beta-binomial test.

4.4.4 Epididymis-to-sperm RNA transfer was not detected

Since a few reports suggest that small RNAs, namely miRNAs and tRFs, synthesised in the epididymal epithelium are packaged into epididymosomes and are delivered to sperm (Chen et al., 2016; Sharma et al., 2016, 2018), this hypothesis was tested with *Spink8-Cre*$^+$ mice that label RNA in the epididymal epithelium.

First, small RNA labelled by 4-thiouracil in the epididymis was confirmed by comparing the labelling level of each small RNA in *Spink8-Cre*$^+$ and *Spink8-Cre*$^-$. Since the previous studies suggest that both miRNAs and tRFs may be mobile, the labelling level of these two small RNA species was assayed. Surprisingly, although a few tRFs are labelled, no significantly labelled miRNAs were detected in the epididymis (Fig. 4.11A, B). Since similar

sets of small RNA were detected in these samples, and the epididymal mRNA was confirmed to be successfully labelled in the previous chapter, this could be due to the low synthesis rate of miRNA in the epididymis. Also, no detectable labelling of both miRNA and tRF in the serum and sperm was observed.

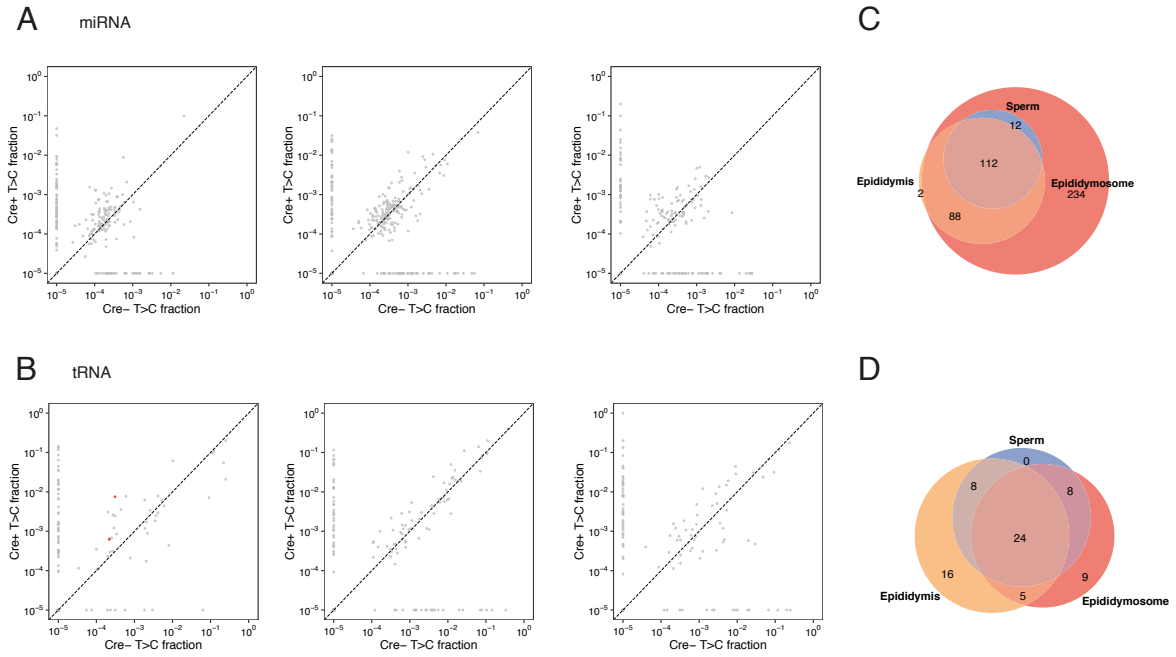


Fig. 4.11 Small RNA mobility from epididymis to sperm was analysed
(A, B) Mean T>C rate among biological replicates (N = 3 each) is compared between *Spink8-Cre*$^{+}$ and *Spink8-Cre*$^{-}$ in epididymis, epididymosomes, and sperm is shown for (A) miRNAs and (B) tRFs. A constant value of 10^{-5} was added when plotting. Red points represent the small RNAs that were significantly labelled in *Spink8-Cre*$^{+}$ mice. (C, D) Euler diagrams showing overlaps of the detected small RNAs among the epididymis, epididymosomes, and sperm for (C) miRNAs and (D) tRFs.

Next, to compare the RNAs detected in three different tissues, Euler diagrams summarising overlaps of the detected small RNAs in these tissues were made (Fig. 4.11C, D). Since the labelled small RNAs were detected in all the samples, the non-detection of the labelled small RNA in epididymosomes and sperm might not be due to the failure in capturing the labelled small RNA species in the small RNA-seq library preparation.

4.5 Discussion

An increasing number of papers have reported mobile RNA in mammalian tissues in the past few decades, due to its promise to develop new diagnostic markers that can be used for an

early detection of various types of diseases based on a signature of extracellular RNA. Also, it has been discussed in the context of epigenetic memory across generations. Similar to the systemic and multigenerational RNAi in *C. elegans*, some reports claim that small RNAs are also mobile from somatic cells to germ cells in *M. musculus*. This field is potentially a very exciting new avenue since if the abundance of RNA delivered from somatic cells to gametes affects the development of embryos derived from it, this could actively relay parental experience to their offspring, which adds another layer of the transgenerational flow of information in addition to the stable DNA sequence.

However, it has been challenging to show the mobility of endogenous RNA directly. Since many identical RNA molecules are synthesised in different cells, it is essential to employ a strategy to distinguish RNAs that are exogenously delivered from those that are natively synthesised to prove the mobility of RNA. In this study, a cell-type-specific metabolic RNA labelling method, SLAM-ITseq, was employed in order to test the intercellular mobility of RNA in three different pairs of tissues, which were chosen based on previous reports showing the mobility of either transgene-derived RNA or extracellular vesicles. However, our analyses with SLAM-ITseq did not detect any small RNA mobility reported in these publications.

Since the absence of evidence is not the evidence of absence, it is not possible to conclude that there is no mobile RNA based on the non-detection of mobile RNA in this study. It could be just due to a lack of detection power: although it was confirmed that this method is sensitive enough to detect 5% of labelled RNA that is diluted in 95% of unlabelled RNA based on the *Tie2-Cre* experiment in brain (Chapter 3), mobile RNA delivered to a recipient cell could be much less abundant than this. Another possibility is that mobile RNA could be very unstable. Our experiment was designed to capture both stable and unstable transcripts, and this is why the mice were culled 6 h after the last injection. However, it is possible that mobile RNA was already degraded within this 6 h. Finally, in this experiment, although the majority of highly expressed RNAs were labelled in the donor cells, not all the donor RNAs were labelled. Thus, the less abundant unlabelled host RNAs may have escaped the detection even though they were mobile.

Nevertheless, this study provides new insights into the nature of endogenous mobile RNA in mammals. As even the most abundant and highly labelled RNAs were not detected in both serum and recipient tissues, RNA release from a cell must be highly selective or very miniature. This is considerably different from the results that were previously reported in experiments using cell lines, where highly expressed miRNAs in the cell lines were also observed in extracellular space. This signifies the importance of performing an experiment *in*

vivo with an appropriate RNA labelling method to assess the significance of various mobile RNA phenomena.

Chapter 5

Detection of zygotic transcription with SLAMseq

5.1 Background

Animal development begins with fertilisation. A haploid sperm and an oocyte fuse together to form a diploid zygote. Just at the time of fertilisation, the majority of RNA contained in zygote is of oocyte-origin because of a massive difference in the cell size between the oocyte (60-70 μm diameter in mouse) and sperm head (6-8 μm length in mouse). After fertilisation, however, this maternally deposited RNA undergoes extensive degradation, and, in turn, the zygotic genome starts to synthesise its own transcripts to shape the transcriptome required for animal development, which is termed zygotic genome activation (ZGA). Although zygotic transcription is known to be essential for development and some key regulators have been discovered, the complete picture about zygotic genome activation is not yet known. Here, literature on mouse preimplantation development and zygotic transcriptional dynamics is summarised.

5.1.1 Preimplantation development of *M. musculus*

After fertilisation, mouse embryos undergo a number of cell divisions, and when they reach the late blastocyst stage, embryos come out from the outer layer, zona pellucida, and implant themselves onto the internal wall of uterus for further development. Embryonic development before the implantation is referred to as “preimplantation development” and has been studied to understand how terminally differentiated germ cells are reprogrammed to become totipotent embryos. Until the 8-cell stage, an embryo undergoes symmetric cleavages and remains totipotent: any one of the cells can give rise to a viable offspring (Chazaud and

Yamanaka, 2016) (Fig. 5.1). In the next cleavage, the cells are asymmetrically divided to generate outer cells and inner cells, which later become trophectoderm and inner cell mass (ICM) cells, respectively. At the blastocyst stage, ICM cells further differentiate into epiblast, which generates fetal cells, and primitive endoderm. The primitive endoderm later produces extra-embryonic tissues such as placenta.

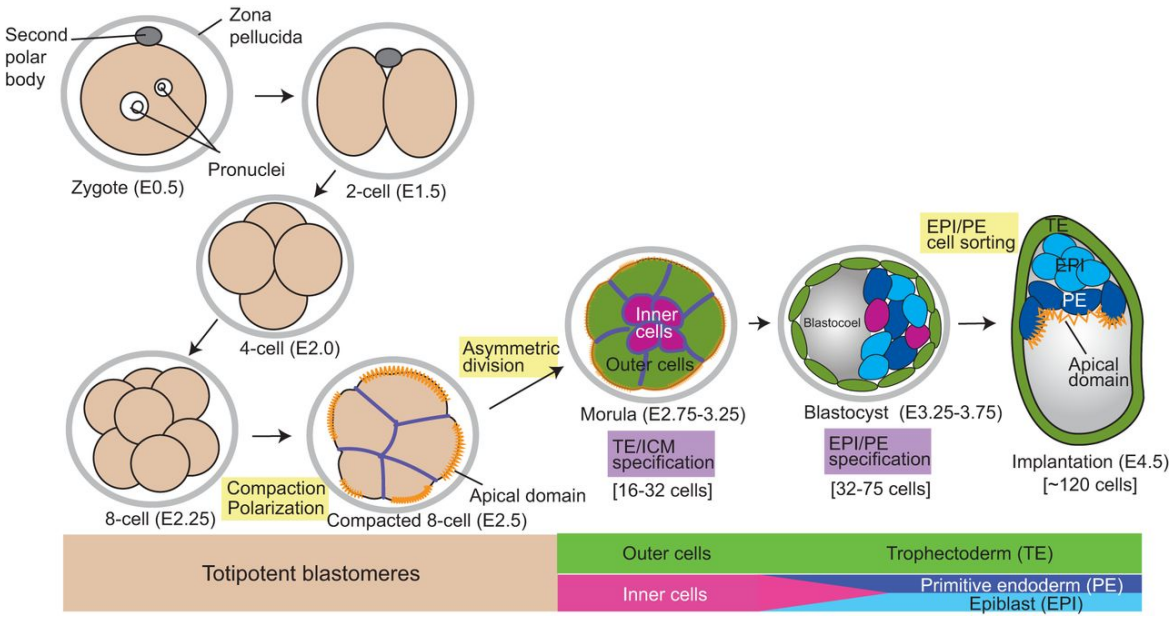


Fig. 5.1 Mouse preimplantation development
Preimplantation development of the mouse embryo is summarised. Figure taken from Chazaud and Yamanaka (2016).

5.1.2 Maternal RNA in preimplantation embryo

Some maternally deposited RNAs are essential for zygotic development. This type of maternal RNA is also called "dormant" RNA since they are not translated in the oocyte but are recruited for translation after fertilisation. This time-specific regulation is achieved through polyA stretching induced later in the developing zygote. This process was shown to be essential for preimplantation development, since the inhibition of polyadenylation by 3′-deoxyadenosine (3′-dA) inhibited genome activation (Aoki et al., 2003). One of the examples of dormant mRNA is the one encoding ORC6L, which is a critical component for DNA replication complex (Murai et al., 2010). Since oocytes between metaphase I (MI) and metaphase II (MII) do not undergo DNA replication but need to replicate after fertilisation to initiate cleavage into 2-cell (2C) embryos, *Orc6l* mRNA is not translated in oocytes but is recruited after fertilisation in embryo.

At the same time, timely degradation of maternal RNA is also critical for proper preimplantation development. A few mechanisms have been proposed to explain how an embryo achieves maternal-specific RNA degradation. Below, the proposed mechanisms are summarised.

miRNA

Although mature oocytes contain little miRNA, late preimplantation embryos contain abundant miRNA, suggesting miRNA is synthesised in the early preimplantation stages. In zebrafish and frog zygotes, miRNAs are shown to target maternal mRNA and accelerate their degradation (Giraldez et al., 2006; Lund et al., 2009). Importantly, the miRNAs that are thought to be involved in the process, namely miR-430 family, are conserved among vertebrates and are expressed in preimplantation embryos.

However, although the same miRNA family is conserved in mice, miRNA does not seem to be essential in mouse preimplantation development. When a miRNA biogenesis factor, *Dgcr8*, is knocked-out in oocytes, they can still generate viable offspring (Suh et al., 2010). Further, even if both the oocyte and sperm lack *Dgcr8*, they can still generate viable offspring. These results suggest that the miRNA pathway is dispensable in mammalian development.

siRNA

Compared with the little phenotypic change observed in *Dgcr8* KO, which is deficient of miRNA maturation, embryos generated from *Dicer* KO oocytes exhibit a dramatic transcriptional change, mitotic defect, and developmental arrest. Since *Dicer* is involved in both siRNA and miRNA processing, these results collectively suggest that endogenous siRNA (endo-siRNA) is essential for preimplantation mouse development, while miRNA is dispensable for the process.

There are a few important properties in mammalian oocytes that allow a robust siRNA machinery. Normally, if dsRNA is present in a mammalian cell, it can induce an inflammatory response through interferon (Kang et al., 2002; Yoneyama et al., 2004). However, the mammalian oocyte lacks an interferon response and induces RNAi through dsRNA injection (Stein et al., 2005). Also, *M. musculus* possesses the oocyte- and *Muridae*-specific Dicer isoform called Dicer^{O} (Flemr et al., 2013). The expression of Dicer^{O} is driven by an intronic insertion of a transposon-like element called MT, and the removal of MT phenocopies *Dicer* KO and *Ago* KO. This suggests that the loss of Dicer^{O}-derived endo-siRNA is responsible for the meiotic defect observed in the *Dicer* KO oocyte.

PolyA tail shortening

Upon fertilisation, some mRNAs undergo polyA tail shortening and are rapidly degraded (Audic et al., 1997). CCR4-NOT complex has been shown to be important in this deadenylation process in the preimplantation embryo (Mishima and Tomari, 2016). Although this process was shown to occur in various cell types and organisms (Subtelny et al., 2014), how such selective mRNA deadenylation is regulated is still unclear.

3′ end uridylation

Recently, non-templated addition of uridine was discovered at the 3′ end of mRNAs targeted by miRNA in various organisms from *Arabidopsis* to mice (Shen and Goodman, 2004). Later studies discovered that terminal uridyltransferases called TUT4 and TUT7 are involved in this non-templated addition of uridine at the 3′ end of RNA (Lim et al., 2014). Analyses of the 3′ end of RNA in the preimplantation embryo revealed that maternal RNA is often uridylated, and the 3′ end uridylation is more frequently observed after shorter polyA tails (Morgan et al., 2017). TUT4/7 knock-down by morpholinos resulted in embryonic defects in *Xenopus laevis*. This suggests that maternal RNAs with short polyA are uridylated by TUT4/7 for degradation, which is important for embryonic development.

5.1.3 Zygotic genome activation (ZGA)

In addition to maternal RNA clearance, transcription from the zygotic genome is also essential for embryonic development. Different organisms seem to employ different machineries to achieve ZGA.

Fly

In *Drosophila*, the key transcription factor that induces ZGA, called Zelda, has been identified (Liang et al., 2008). This zinc-finger protein binds to a *cis*-regulatory heptamer motif, CAGGTAG, which is shared among the majority of zygotic genes. Zelda-lacking embryos cannot induce zygotic genome activation and are defective in development. Thus, Zelda controls the expression of these zygotic genes, which are essential for embryonic development.

Mammals

Nascent transcription labelling assay with BrU revealed that zygotic transcription happens as early as the late 1-cell stage (Bouniol et al., 1995). Since the most transcripts in this stage are not properly spliced (Abe et al., 2015), it was questioned if the transcripts synthesised

in the 1-cell zygote are functional. However, inhibition of these early transcripts led to 2C arrests, suggesting that this transcription is essential for development (Abe et al., 2018). The highest rate of zygotic transcription was observed in the 2C embryos, and transcriptional inhibition in this stage led to developmental arrest, suggesting that transcription at the 2C stage is necessary for embryonic development. However, the transcriptional cascade that governs zygotic genome activation in mammals is still poorly understood.

Three papers claiming that a transcription factor, Dux, governs zygotic genome activation in human and mouse embryos were published in the same issue of *Nature Genetics* (De Iaco et al., 2017; Hendrickson et al., 2017; Whiddon et al., 2017). However, unlike Zelda in *Drosophila*, following works reported that Dux-lacking mouse embryo can develop in term and generates viable mice (Chen and Zhang, 2019; De Iaco et al., 2019), though lower developmental potential was observed. These studies suggest that Dux is an important factor affecting a significant proportion of mammalian zygotic transcripts but not essential for development.

5.1.4 Reactivation of transposable elements during ZGA

Another notable phenomenon during ZGA is that a number of different classes of transposable elements (TEs) are derepressed. TEs are selfish genetic elements that exist in various organisms and account for more than 30% of the mouse genome. There are various types of TEs, and each has a distinct mobilisation machinery (Fig. 5.2). To maintain the genomic integrity, their expression is tightly regulated by multiple molecular machineries such as DNA methylation, PIWI-interacting RNA (piRNA) (Aravin et al., 2007), and KRAB-zinc finger proteins (Imbeault et al., 2017). In mammalian embryos, however, it is known that transcripts derived from a number of different TEs are detected in multiple stages of embryos (Evsikov et al., 2004; Kigami et al., 2003; Peaston et al., 2004). Especially, abundant transcripts derived from endogenous retrovirus (ERV) are observed in 2C mouse embryo and, notably, MuERV-L (murine endogenous retrovirus with leucine tRNA primer) is well-known to be expressed only during the 2C stage (Kigami et al., 2003). Surprisingly, many 2C-activated genes were discovered to have a MuERV-L-derived promoter (Macfarlan et al., 2012), suggesting that TE integrations into intergenic regions contribute to forming the 2C transcriptional network. Also, there is a subpopulation of ES cells that expresses MuERV-L, and this population exhibits 2C-like transcriptome, suggesting that MuERV-L expression is linked with the 2C gene regulatory network (Macfarlan et al., 2012).

In addition to these co-opted roles of TE-derived sequences in the genome, it is also speculated that transposon-derived RNA itself may have biological roles in the host cell. LINE-1 RNA was shown to recruit Kap1 and Nucleolin to rDNA and Dux-encoding loci, resulting in

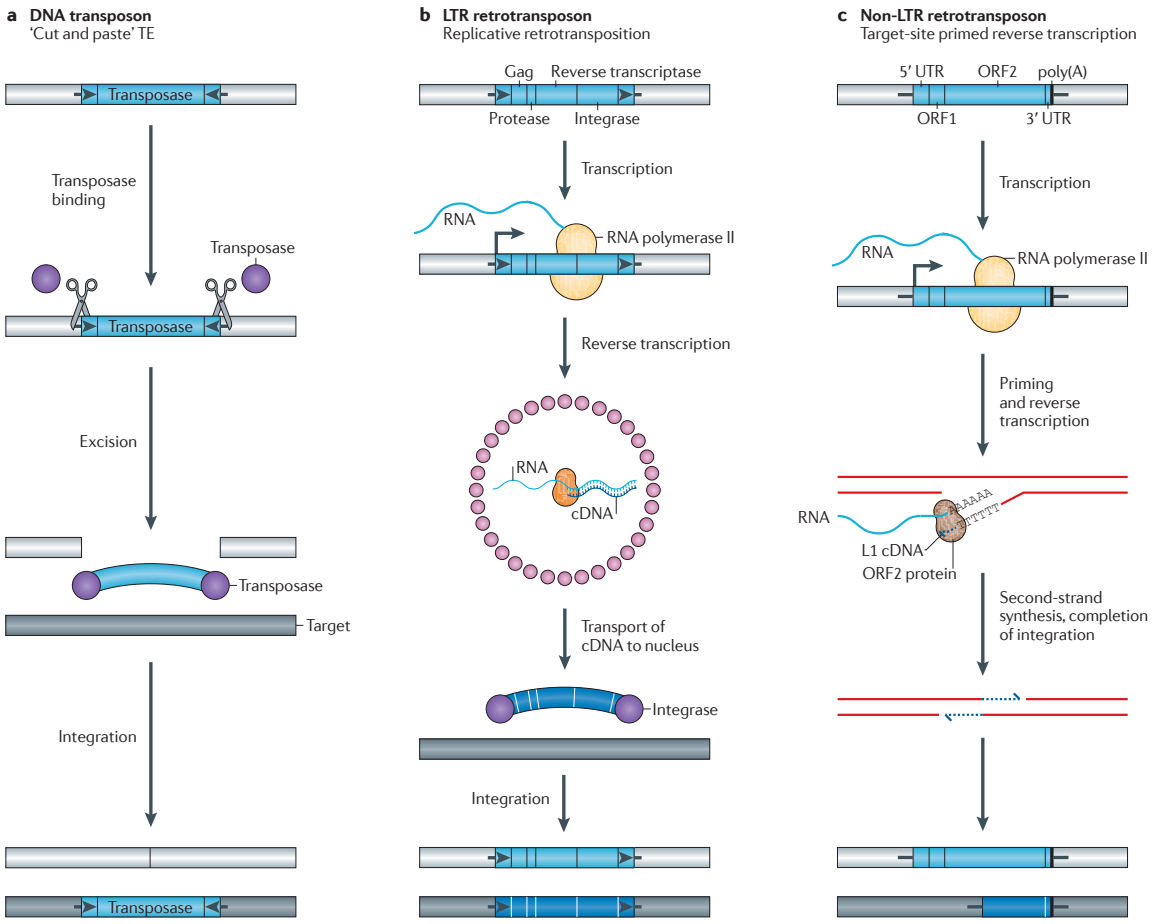


Fig. 5.2 Mobilisation mechanisms of different TE families
(a) DNA transposons excise their own sequence using a transposase encoded and integrate into a new genomic locus. Retrotransposons (b, c) employ a "copy and paste" mechanism through generating RNA intermediate. (b) RNA synthesised from an LTR transposon is first transferred to the cytoplasm and reverse-transcribed to synthesise cDNA in a virus-like capsid made of Gag proteins (pink circle). The cDNA is then transferred back to the nucleus and integrated into a new genetic locus. (c) Non-LTR transposons use an endonuclease to induce a single-stranded nick at an integration site of the genomic DNA and reverse transcribe the TE transcript at the integration site for integration. Taken from Levin and Moran (2011).

the repression of Dux and induction of rRNA synthesis in the mouse blastocyst (Percharde et al., 2018). Although the precise mechanisms of how LINE-1 RNA targets different genetic loci and induces the opposite transcriptional responses are still unknown, this study shows that retrotransposon-derived RNA could function as a regulatory RNA.

5.1.5 Experimental approaches to study ZGA

Since first discovered in 1964 (Mintz, 1964), transcriptional activity in the early preimplantation embryos has been widely studied using various methods. Here, different approaches to study zygotic transcription are summarised.

Metabolic labelling with uridine analogues

Radiolabelled uridine, namely [^{3}H]-uridine was the first tool to dissect ZGA. Embryos at different stages were exposed to the analogue, and the RNA synthesis rate was measured (Mintz, 1964). It is surprising that this first experiment even revealed that the RNA synthesised in the 2C embryo mostly remains in the nucleus. Also, combined with size fractionation with electrophoresis and polyA selection, it was discovered that polyA RNA is also actively synthesised in the 2C embryos (Clegg and Pikó, 1983a; Ellem and Gwatkin, 1968). It is particularly worth noting that RNA synthesis at the 1-cell stage (i.e. zygote) was detected with the metabolic labelling method (Clegg and Pikó, 1983b). Also, extensive polyadenylation for new RNA synthesis and pre-existing RNA degradation were observed.

As an alternative to the radiolabelling methods, 5-bromouridine-5′-triphosphate (BrUTP) was used to capture the first wave of ZGA with higher sensitivity (Bouniol et al., 1995). They not only unambiguously determined that the first ZGA occurs at the late 1-cell stage, but also identified that the first transcripts are synthesised in the paternal pronucleus.

Transcriptional perturbation

Another key strategy in studying ZGA is to block the transcription by stalling polymerases. The most frequently used agent is α-amanitin, which specifically binds to Pol II and stalls it. Pol II-dependent transcription in the 2C embryo was shown with decreased expression of [^{3}H]-uridine labelled RNA in α-amanitin-treated embryos (Warner and Versteegh, 1974).

While the transcriptional inhibition by α-amanitin is irreversible, inhibition with 5,6-dichloro-1-β-D-ribofuranosyl-benzimidazol (DRB) is reversible (Tamm et al., 1976), and thus DRB has been used to study the significance of transcripts synthesised at a particular stage of embryo by assessing developmental progression after a stage-specific treatment (Abe

et al., 2018). This study identified that the minor transcription observed in the late 1-cell stage is essential for later zygotic development.

Comparison of RNA abundance

Since the methods above do not provide information about abundance of particular RNA molecules, other experimental approaches that provide transcript-level information have also been used to study the transcriptional dynamics of particular genes. Northern blot and RT-qPCR have been widely used to achieve low-throughput quantification of transcripts in different stages of embryos. Especially, by comparing the embryonic transcriptome with that of oocytes, transcripts synthesised in the zygote can be identified. Inventions of high-throughput methods, such as microarray or RNA-seq, enabled to simultaneously study the dynamics of a diverse set of transcripts during ZGA (Hamatani et al., 2004; Tang et al., 2009).

Also, these quantitative methods have been combined with the above two methods. Labelled RNA from 4-thiouridine-exposed embryos was isolated with biotinylation combined with the streptavidin beads pull-down, and the pulled-down RNA was quantified to identify transcripts synthesised in the embryo (Heyn et al., 2014). Hamatani et al. combined α-amantin treatment with microarrays and identified genes actively synthesised in different stages of embryos.

SNP assay

Since RNA-seq determines the exact sequence of each transcript, single-nucleotide polymorphism (SNP) in each transcript can also be detected. Considering that the abundance of sperm-deposited RNA is negligible, the zygotic transcriptome just after fertilisation contains SNPs only from the maternal allele. Thus, by identifying transcripts with paternal SNPs after ZGA, zygotic transcripts can be identified (Harvey et al., 2013; Lee et al., 2013). Also, with the development of single-cell RNA-seq method (Tang et al., 2009), this strategy can also be applied to analyse SNP using non-inbred animals including humans (Xue et al., 2013). However, this method is not exhaustive in identifying all the zygotic transcripts because informative SNPs to differentiate the alleles are present up to 20% of all the detected transcripts, and many genes exhibit the maternal monoallelic expression pattern. Still, it is a useful method to study ZGA in genetically diverse samples and to study the allelic expression patterns.

5.2 Aims of this chapter

The most simple strategy to study ZGA is to compare the transcriptome between the oocyte and the zygote/2C embryos. Although this method has been widely used and has successfully identified a number of transcripts synthesised in the 2C embryos (Hamatani et al., 2004), this approach has some limitations. First, since this method just focuses on the abundance of transcripts for each stage of embryos, it is difficult to identify transcripts that are synthesised both in the oocyte and the embryo. It is also very challenging to compare the abundance of RNA between different stages of embryos since most RNA-seq pipelines are optimised to compare cells with more or less similar abundance of total RNA and proportions of RNA species (Anders and Huber, 2010; Robinson et al., 2010). In the oocyte-to-embryo transition process, however, maternal RNA degradation and ZGA alter both the proportion of RNA species and the total abundance of RNA. Although one possible solution is to include a spike-in, it is also difficult to find the right sample-to-spike-in ratio.

With metabolic RNA labelling, these problems can be solved. By culturing embryos in nucleotide analogue-containing medium, only RNA synthesised in embryos incorporates the analogue. Also, by restricting the exposure time so that embryos at a particular stage are labelled, RNA synthesised in that particular stage can specifically be labelled, which is powerful in studying stage-specific active gene expression, as opposed to RNA abundance in each embryonic stage.

In this chapter, mouse 2C transcriptome is studied with SLAMseq. Mouse 2C embryos were cultured *in vitro* in medium containing 4-thiouridine, and the RNA isolated from the embryos was analysed to find actively synthesised transcripts in the 2C embryo.

5.3 Optimisation of RNA labelling condition

Since 4-thiouridine exposure has been reported to inhibit cell proliferation through blocking rRNA synthesis (Burger et al., 2013), we first confirmed the highest concentration of 4-thiouridine that can be used to culture the mouse embryos without affecting the preimplantation development. WT embryos were cultured in medium with different 4-thiouridine concentrations. As shown in Table 5.1, the embryos cultured with 0, 1, and 5 mM of 4-thiouridine developed to the blastocyst stage at almost identical rates, while none of the embryos in 10 mM 4-thiouridine reached the blastocyst stage. Thus, in the subsequent analyses, the embryos were cultured with 5 mM 4-thiouridine to achieve the labelling without any developmental defects.

Table 5.1 Proportions of the embryos that reached the blastocyst stage with different 4-thiouridine concentrations

4-thiouridine conc. (mM)	Number of starting zygotes	Number of blastocysts obtained
10	15	0 (0%)
5	15	15 (100%)
1	15	14 (93%)
0	10	10 (100%)

5.4 Analysis of active transcription in the 2C embryo

5.4.1 Higher T>C was observed in 4-thiouridine-exposed embryos

To test if this method can label RNA synthesised during ZGA, the mouse embryos were collected and *in vitro* cultured for 24 h until the embryos reached the 2C stage. The 2C embryos were then transferred to medium containing 5 mM of 4-thiouridine or control medium and were further incubated for 4 h. RNA was extracted from the embryos and used as input for SLAMseq. By comparing the T>C rate between 4-thiouridine-exposed embryos and control embryos, transcripts with a significantly higher T>C conversion rate in 4-thiouridine embryos compared with control were identified (Fig. 5.3A). To see if 4-thiouridine exposure affects the transcriptome of embryos, the abundance of RNA was compared. As shown in Fig. 5.3B, a highly-positive correlation was observed between the abundance of transcripts in control and 4-thiouridine embryos, which suggests that 4-thiouridine incorporation did not affect RNA expression levels in the embryos.

5.4.2 SLAMseq labelled the zygotic transcripts specifically

The labelled transcripts were further analysed for characterisation. To confirm that the labelled transcripts are in agreement with the previously identified zygotic transcripts, the obtained list of labelled genes was compared against the previously published 2C gene list (Macfarlan et al., 2012) (Fig. 5.4). Although the number of genes in the two lists are different, potentially due to SLAMseq being performed with only 4 h exposure, more than 40% of SLAMseq-identified genes overlap with the 2C genes in the dataset.

Also, to confirm the labelling level of the maternally-deposited RNA in the 2C embryo, a list of genes that are highly expressed in metaphase II (MII) oocytes was obtained from a previous literature (Tang et al., 2009), and this confirmed the labelling level of these genes in the 2C (Fig. 5.5). As expected, although the oocyte-synthesised transcripts are of high abundance in the 2C embryo, the majority of them are not labelled. It is interesting that there

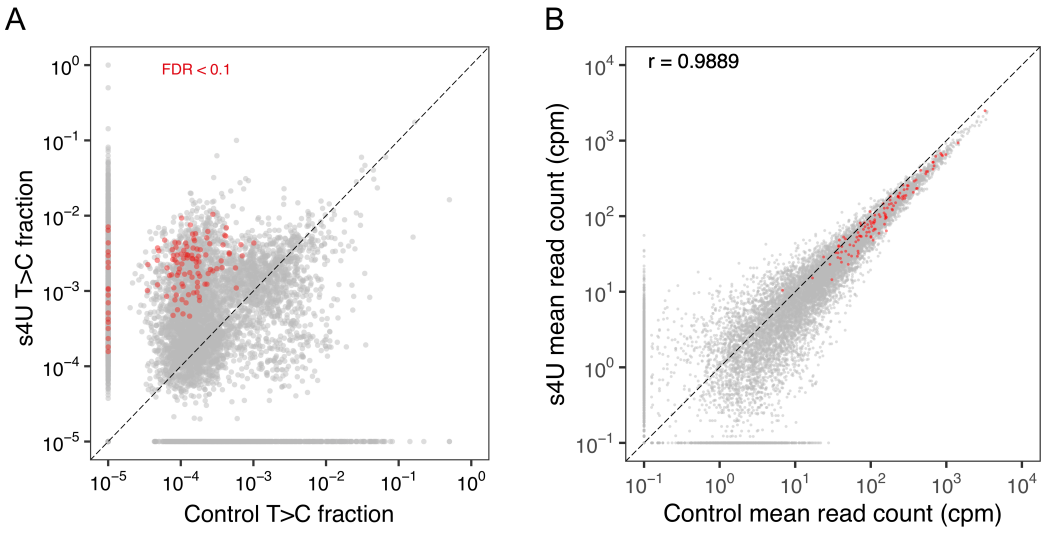


Fig. 5.3 SLAMseq identified genes actively transcribed in the 2C embryos
(A) Mean T>C conversion rates among biological replicates (N = 3 each) for each polyA transcript in the control and 4-thiouridine embryos are shown. Genes with a significantly (FDR <0.1, beta-binomial test) higher T>C rate in 4-thiouridine are shown in red. A constant value of 10^{-5} was added to each value when plotting. (B) Mean gene expression in 4-thiouridine and control embryos is shown. Spearmen's correlation coefficient is shown. A constant value of 1 was added to each value when plotting.

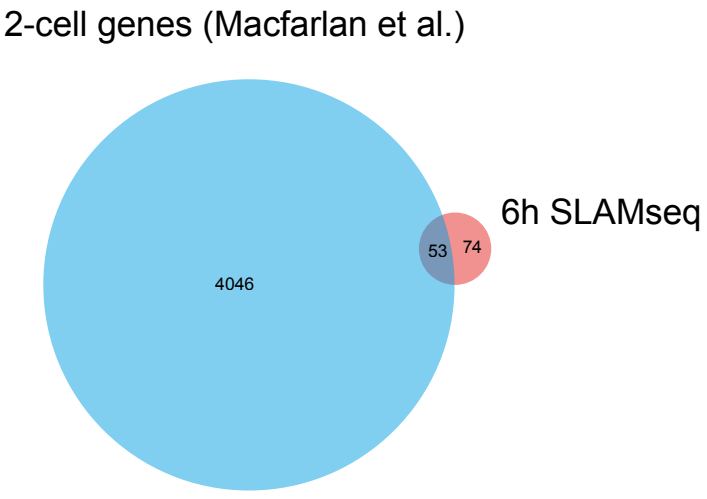


Fig. 5.4 Venn diagram comparing the labelled genes and the 2C genes
Genes labelled with SLAMseq in the 2C embryos were compared with the genes identified to be 2C-expressed in a previous report (Macfarlan et al., 2012) and shown as a Venn diagram.

are several oocyte genes that are labelled, as this suggests that SLAMseq could potentially be able to identify genes that are transcribed in both oocyte and zygote, which has been challenging to achieve with conventional RNA-seq methods.

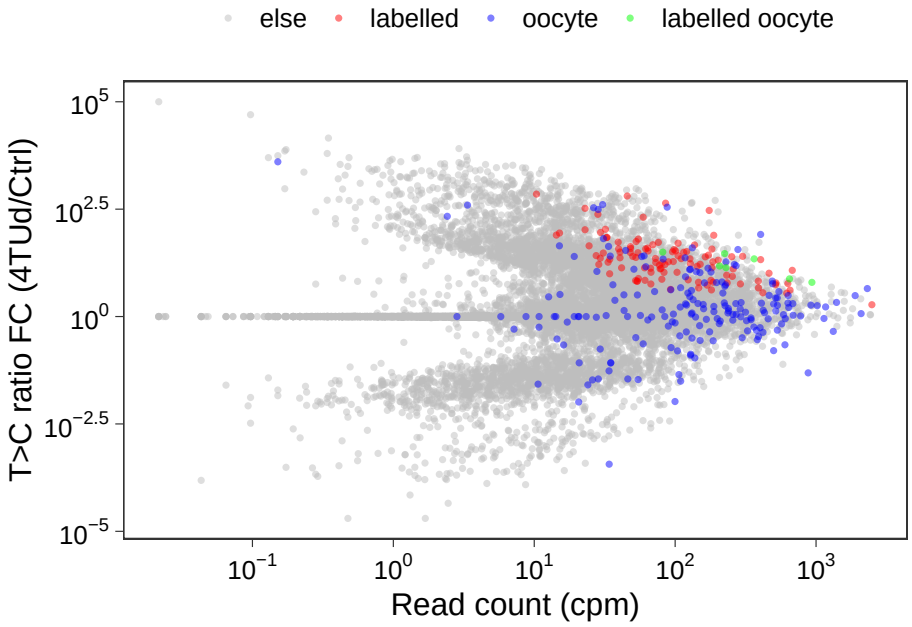


Fig. 5.5 Abundance and labelling level of the transcripts detected in the 2C embryo
Read count and fold change of labelling level of the 4-thiouridine embryo divided by the control embryo are shown. To highlight the maternally-deposited RNA, a list of genes that are highly expressed in MII oocytes (oocyte genes) are obtained from Tang et al. (2009). Unlabelled oocyte genes, blue; labelled oocyte genes, green; labelled non-oocyte genes, red.

5.4.3 No significantly enriched DNA motif was found upstream of the labelled genes

To dissect the molecular mechanisms that govern the expression of genes in the early 2C embryo, common DNA sequence motifs upstream of the labelled genes were sought using HOMER (Heinz et al., 2010), and the enrichment levels of transcription factor (TF) binding sites were quantified (Fig. 5.6). The analysis has revealed that a number of known TF-binding motifs are enriched among the upstream of the labelled transcripts identified by SLAMseq; however, all of them have a FDR >0.1, suggesting that no strong enrichment was observed.

Total Target Sequences = 121, Total Background Sequences = 26537

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Backgroun Sequences with Motif
1		Fli1(ETS)/CD8-FLI-ChIP-Seq(GSE20898)/Homer	1e-3	-7.945e+00	0.1518	64.0	52.89%	9898.2	37.30%
2		ETV1(ETS)/GIST48-ETV1-ChIP-Seq(GSE22441)/Homer	1e-3	-7.616e+00	0.1518	63.0	52.07%	9794.1	36.91%
3		ETS1(ETS)/Jurkat-ETS1-ChIP-Seq(GSE17954)/Homer	1e-3	-7.421e+00	0.1518	53.0	43.80%	7824.2	29.48%
4		NRF(NRF)/Promoter/Homer	1e-3	-7.277e+00	0.1518	36.0	29.75%	4651.4	17.53%
5		GABPA(ETS)/Jurkat-GABPa-ChIP-Seq(GSE17954)/Homer	1e-2	-6.651e+00	0.1518	53.0	43.80%	8066.5	30.40%
6		PU.1(ETS)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-2	-6.511e+00	0.1518	21.0	17.36%	2273.5	8.57%
7		ELF1(ETS)/Jurkat-ELF1-ChIP-Seq(SRA014231)/Homer	1e-2	-5.705e+00	0.2035	49.0	40.50%	7586.2	28.59%
8		Lhx3(Homeobox)/Neuron-Lhx3-ChIP-Seq(GSE31456)/Homer	1e-2	-5.672e+00	0.2035	29.0	23.97%	3810.5	14.36%
9		Unknown(Homeobox)/Limb-p300-ChIP-Seq/Homer	1e-2	-5.532e+00	0.2035	14.0	11.57%	1365.3	5.14%
10		Elk1(ETS)/Hela-Elk1-ChIP-Seq(GSE31477)/Homer	1e-2	-5.513e+00	0.2035	52.0	42.98%	8258.2	31.12%
11		ERG(ETS)/VCaP-ERG-ChIP-Seq(GSE14097)/Homer	1e-2	-5.327e+00	0.2035	56.0	46.28%	9147.6	34.47%
12		Elf4(ETS)/BMDM-Elf4-ChIP-Seq(GSE88699)/Homer	1e-2	-5.225e+00	0.2035	49.0	40.50%	7759.9	29.24%
13		ETV4(ETS)/HepG2-ETV4-ChIP-Seq(ENCODE)/Homer	1e-2	-5.168e+00	0.2035	60.0	49.59%	10044.4	37.85%
14		KLF3(Zf)/MEF-Klf3-ChIP-Seq(GSE44748)/Homer	1e-2	-4.905e+00	0.2264	54.0	44.63%	8901.2	33.54%
15		SPDEF(ETS)/VCaP-SPDEF-ChIP-Seq(SRA014231)/Homer	1e-2	-4.891e+00	0.2264	37.0	30.58%	5523.5	20.81%
16		Elk4(ETS)/Hela-Elk4-ChIP-Seq(GSE31477)/Homer	1e-2	-4.838e+00	0.2264	51.0	42.15%	8313.6	31.33%
17		ETS(ETS)/Promoter/Homer	1e-2	-4.785e+00	0.2264	37.0	30.58%	5559.7	20.95%

Fig. 5.6 Enriched sequence motifs upstream of the 2C-labelled genes
Common DNA sequence motifs from -300 to +50 relative to transcription start site (TSS) of the genes labelled in the 2C embryo were sought with HOMER. Enrichment level of known TF-binding sequences are summarised.

5.5 Analyses of active TE transcription in the 2C embryo

5.5.1 Labelled TE transcripts were identified with SLAMseq

Since TE-derived transcripts are known to be expressed in the 2C mouse embryo, we further analysed the SLAMseq data to see if TE transcripts were also labelled with 4-thiouridine exposure. By comparing the T>C rate between the 4-thiouridine and control embryos, significantly labelled TE transcripts were discovered. Since there are multiple identical copies of each TE in the genome, it is often not possible to align a read to a single TE-encoding genomic locus unambiguously. Thus, if a read cannot be aligned to a unique genetic locus, the read is mapped to a locus randomly among all the candidate loci, and, subsequently, the read count as well as T>C conversion events were summarised at the TE gene level.

The beta-binomial test on the T>C count data for each TE gene identified the TE genes that have higher T>C rates in the 4-thiouridine embryos than the control embryos (Fig. 5.7A). Also, the abundance of each TE transcript was compared between 4-thiouridine and control embryos, and a high positive correlation (Pearson's correlation coefficient = 1.00) was observed (Fig. 5.7B). These results suggest that the actively synthesised TE-derived transcripts were labelled with the 4-thiouridine exposure, and the transcription of TEs was not affected by 4-thiouridine incorporations.

To better characterise the labelled TE transcripts, all the expressed TEs were sorted by expression and plotted (Fig. 5.8). The majority of highly abundant TE transcripts in the early 2C embryo were confirmed to be labelled with SLAMseq. Also, these labelled TE genes include the well-known 2C-synthesised TEs, such as MuERV-L. These results suggest that the majority of highly abundant TE transcripts in the 2C embryo are synthesised in this particular stage and not inherited from the oocyte.

5.5.2 Different classes of TE genes are active in the early 2C

Further, to summarise the labelled TEs by their classes, the number of labelled TE genes in each TE class was compared between the labelled and unlabelled fractions (Fig. 5.9). Many ERVs and LINE transposons are significantly labelled, which is in agreement with previous reports (Fadloun et al., 2013; Macfarlan et al., 2012). Intriguingly, when compared to the proportions of the unlabelled TE genes, SINEs are overrepresented in the labelled TE genes. This result suggests that the majority of SINE-derived RNAs detected in the 2C embryos are synthesised in the 2C stage.

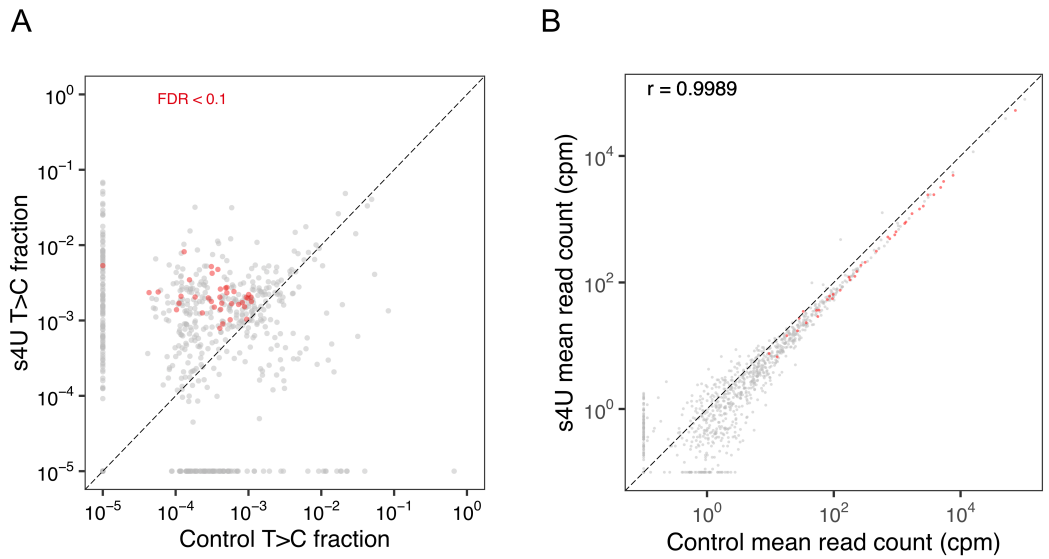


Fig. 5.7 Actively transcribed TE-derived RNAs were identified in the 2C embryos (A) Mean T>C rate of each TE gene in the 4-thiouridine and control embryos (*N* =3 each). TE genes with a significantly (FDR <0.1) higher T>C rate in the 4-thiouridine embryos are shown in red. A constant value of 10^{-5} was added to each value when plotting. (B) Expression of each TE-derived RNA in the 4-thiouridine and control embryos. Spearman's correlation coefficient is shown. A constant value of 1 was added to each value when plotting.

5.6 Discussion

Capturing active transcription is essential in studying the transcriptional network in the cell. In this chapter, I showed that 4-thiouridine treatment followed by SLAMseq identifies active transcription of polyA and TE-derived transcripts and distinguishes zygotic transcripts from maternally-deposited RNA. The concentration of 4-thiouridine was first optimised to maintain the viability and developmental potential of the embryos. Also, it was confirmed that 4-thiouridine exposure at this concentration did not affect the abundance of each transcript, and thus it suggests that 4-thiouridine exposure itself has little effect on RNA metabolism in embryos, and that the method captures the native transcriptional state of the embryo.

Conventionally, zygotic transcripts have been studied by comparing RNA-seq data obtained from different stages of the embryos to identify transcripts that are significantly more abundant in a particular stage than the previous stage. This approach has been successful in identifying zygotic transcripts that are not synthesised in the oocytes. However, to understand the transcriptional cascade that governs the ZGA, it is essential to unbiasedly capture genes that are transcriptionally active at a particular stage of embryos. SLAMseq is the first method that analyses transcriptional activity with the nucleoside incorporation rate at single-base resolution. Application of this method at different stages of the embryos would reveal a

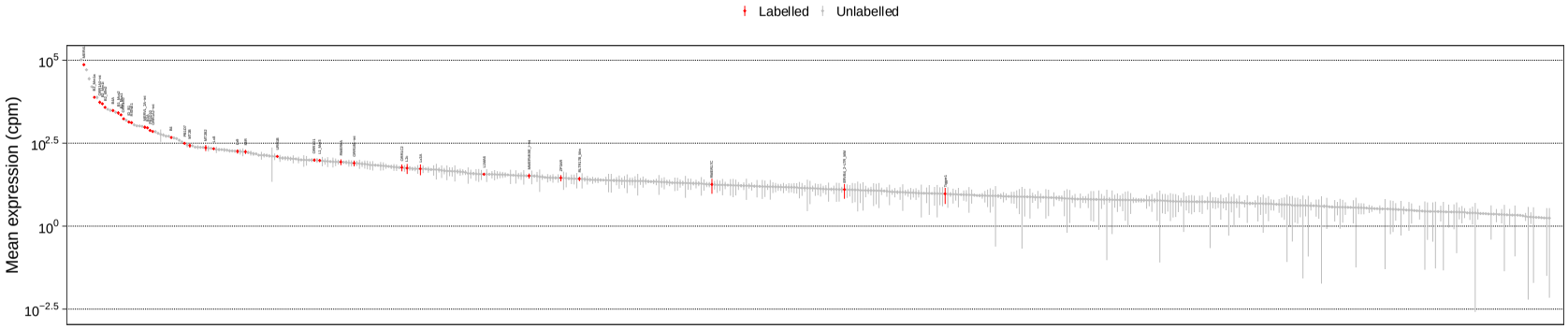


Fig. 5.8 TE transcripts sorted by expression
TE genes detected in all the embryo samples are sorted by their mean expression (N = 3 each) and shown in descending order. Diamonds, mean expression; red, labelled genes; vertical lines, 95% confidence intervals.

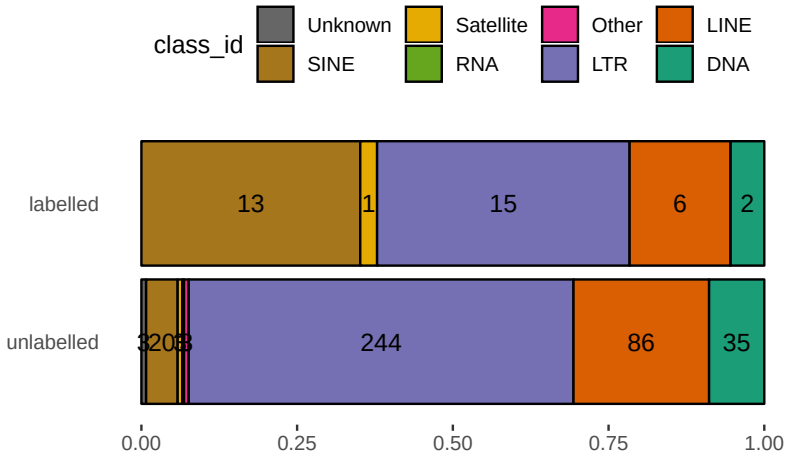


Fig. 5.9 Proportions of the labelled TE classes in the 2C embryos
The fractions of different TE families in the labelled and unlabelled TE genes are summarised. The numbers shown in each bar plot represent the number of TE genes.

complete transcriptional map at each stage of the embryos. Further, by combining molecular perturbations (e.g. inhibition of a transcription factor), transcripts directly under control of the particular factor can be studied.

SLAMseq also has a practical benefit compared to existing metabolic RNA labelling methods. Conventionally, metabolically labelled transcripts are isolated from the pool of RNA by a biochemical method, and the isolated RNA fraction as well as the input fraction are sequenced separately. In addition to the analytical difficulties with this approach discussed in Chapter 3, it requires abundant starting materials to obtain sufficient pulled-down RNA for sequencing library construction, which is particularly difficult to achieve with embryo experiments, as more mice would have to be used for superovulation. With SLAMseq, labelled and unlabelled RNA are sequencing in the same library; therefore, the minimum number of embryos sufficient for a library is needed for each biological replicate, though more embryos may be needed if transcriptionally less active genes have to be captured.

In this study, SLAMseq was performed to label both cells of the 2C embryo, as these two cells are considered to be identical. However, by generating mice expressing UPRT in a particular cell population, SLAM-ITseq could be used to study transcriptional activity in a specific subpopulation of cells in later embryos, where cells are differentiated and exhibit different transcriptome (e.g. blastocyst).

Further experiments are needed to determine whether transcriptional activity in different stages of the embryos can be studied with this metabolic RNA labelling method. Since we only confirmed that SLAMseq labels RNA transcribed in the 2C stage, where the transcrip-

tional activity is thought to be the highest during the preimplantation development, only little RNA labelling may be achieved in other stages, in which transcriptional activity is less strong.

Chapter 6

Final considerations and future perspectives

Metabolic RNA labelling has led to key biological discoveries. Radiolabelled nucleoside analogues allowed scientists observe short-lived mRNAs in nucleus for the first time (Astrachan and Volkin, 1958; Hershey et al., 1953), which led to the discovery of the flow of genetic information to protein synthesis. Zygotic genome activation as a phenomenon was also first observed with RNA radiolabelling (Mintz, 1964).

Although metabolic labelling has become less common in studying RNA transcriptional changes after the inventions of PCR and various high-throughput approaches, it is still an important tool in interrogating transcriptional dynamics, and, indeed, has discovered miRNA dynamics recently (Duffy et al., 2015).

In this thesis, I showed that SLAM-ITseq can be used to identify cell-type-specific transcriptome without cell sorting. While this method is already a useful method in studying native transcriptional states of particular cells without cell sorting, it would potentially be able to observe new biological phenomena if this method is applied to analyse transcriptional dynamics *in vivo* by combining it with a perturbation. For example, it would be interesting to observe the transient transcriptional changes in gut epithelium after a virus challenge.

Also, SLAM-ITseq enabled to study the intercellular mobility of endogenous RNA for the first time. Although no signs of mobility was detected in the three cell types tested in this thesis, different cell types or cells under different conditions may release mobile RNA in mammals, and SLAM-ITseq might be useful in directly proving the mobility of such RNAs.

SLAMseq was also shown to be useful to capture the active transcription in the mouse 2C embryo. Although various metabolic labelling methods have been used to study ZGA, SLAMseq took the method to the next level. The pull-down-independent approach to identify the labelled RNA allows an unbiased detection of the zygotic transcripts starting from a small

number of embryos. Also, the labelling level is quantified at single-nucleotide resolution, which enables a quantitative analysis of the transcriptional dynamics.

One of the biggest limitations of SLAM-ITseq is the potential sensitivity problem, as the T>C conversion rate that can be achieved is still lower than what can be achieved with 4-thiouridine exposure. To further improve the method, it might be worth testing different UPRTs from various species to discover UPRT with the highest enzymatic activity. With this optimisation, more sensitive data could potentially be obtained from SLAM-ITseq experiments. Also, sequencing strategy could be improved. Since the sequencing error rate inherent to the illumina method could potentially be higher than the T>C rate induced by the labelling, the use of paired-end sequencing or UMI might be effective in lowering the background noise, leading to improving the sensitivity.

The development of SLAMseq and equivalent methods to identify 4-thiouridine incorporations by base-conversions has potential to make metabolic labelling to be applied in wider research fields and to enable more detailed studies of RNA transcription, as opposed to just comparing steady-state abundance of RNA. One of the most promising examples is a study that combined SLAMseq with a transcription factor knock down and identified the genes directly controlled under the transcription factor (Muhar et al., 2018). This clearly overcame the problem in the conventional RNA-seq method that cannot differentiate direct and indirect effects of transcription factor perturbation.

SLAMseq was also combined with a single-cell RNA sequencing (scRNA-seq) method and captured transcriptional dynamics at single-cell level (Erhard et al., 2019). Hence, SLAM-ITseq could potentially be combined with scRNA-seq to capture the heterogeneity of transcriptional dynamics within the same cell type *in vivo*. Also, using F1 hybrid mice, allelic expression analysis could also be performed. By finding transcripts that contain both allele-specific SNPs and T>Cs, one can estimate the allele-level expression for each gene, which would be powerful in studying genomic imprinting or random monoallelic expression.

Metabolic RNA methods would potentially shed light on the still enigmatic eukaryotic transcriptional machinery.

References

Abe, K.-I., Funaya, S., Tsukioka, D., Kawamura, M., Suzuki, Y., Suzuki, M. G., Schultz, R. M., and Aoki, F. (2018). Minor zygotic gene activation is essential for mouse preimplantation development. *Proc. Natl. Acad. Sci. U. S. A.*, 115(29):E6780–E6788.

Abe, K.-I., Yamamoto, R., Franke, V., Cao, M., Suzuki, Y., Suzuki, M. G., Vlahovicek, K., Svoboda, P., Schultz, R. M., and Aoki, F. (2015). The first murine zygotic transcription is promiscuous and uncoupled from splicing and $3'$ processing. *EMBO J.*, 34(11):1523–1537.

Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, 4.

Akay, A., Jordan, D., Navarro, I. C., Wrzesinski, T., Ponting, C. P., Miska, E. A., and Haerty, W. (2019). Identification of functional long non-coding RNAs in C. elegans. *BMC Biol.*, 17(1):14.

Amândio, A. R., Necsulea, A., Joye, E., Mascrez, B., and Duboule, D. (2016). Hotair is dispensible for mouse development. *PLoS Genet.*, 12(12):e1006232.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106.

Aoki, F., Hara, K. T., and Schultz, R. M. (2003). Acquisition of transcriptional competence in the 1-cell mouse embryo: requirement for recruitment of maternal mRNAs. *Mol. Reprod. Dev.*, 64(3):270–274.

Aravin, A. A., Hannon, G. J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, 318(5851):761–764.

Arroyo, J. D., Chevillet, J. R., Kroh, E. M., Ruf, I. K., Pritchard, C. C., Gibson, D. F., Mitchell, P. S., Bennett, C. F., Pogosova-Agadjanyan, E. L., Stirewalt, D. L., Tait, J. F., and Tewari, M. (2011). Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc. Natl. Acad. Sci. U. S. A.*, 108(12):5003–5008.

Astrachan, L. and Volkin, E. (1958). Properties of ribonucleic acid turnover in t2-infected escherichia coli. *Biochim. Biophys. Acta*, 29(3):536–544.

Audic, Y., Omilli, F., and Osborne, H. B. (1997). Postfertilization deadenylation of mRNAs in Xenopus laevis embryos is sufficient to cause their degradation at the blastula stage. *Mol. Cell. Biol.*, 17(1):209–218.

Bagijn, M. P., Goldstein, L. D., Sapetschnig, A., Weick, E.-M., Bouasker, S., Lehrbach, N. J., Simard, M. J., and Miska, E. a. (2012). Function, targets, and evolution of Caenorhabditis elegans piRNAs. *Science*, 337(6094):574–578.

Batista, P. J., Ruby, J. G., Claycomb, J. M., Chiang, R., Fahlgren, N., Kasschau, K. D., Chaves, D. A., Gu, W., Vasale, J. J., Duan, S., Conte, D., Luo, S., Schroth, G. P., Carrington, J. C., Bartel, D. P., and Mello, C. C. (2008). PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Mol. Cell*, 31(1):67–78.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Bouniol, C., Nguyen, E., and Debey, P. (1995). Endogenous transcription occurs at the 1-cell stage in the mouse embryo. *Exp. Cell Res.*, 218(1):57–62.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(5):525–527.

Brennecke, J., Aravin, A. a., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small RNA-Generating loci as master regulators of transposon activity in drosophila. *Cell*, 128(6):1089–1103.

Brenner, S. (1974). The genetics of Caenorhabditis elegans. *Genetics*, 77(1):71–94.

Brenner, S., Jacob, F., and Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581.

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S., and Rastan, S. (1992). The product of the mouse xist gene is a 15 kb inactive x-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, 71(3):515–526.

Burger, K., Mühl, B., Kellner, M., Rohrmoser, M., Gruber-Eber, A., Windhager, L., Friedel, C. C., Dölken, L., and Eick, D. (2013). 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol.*, 10(10):1623–1630.

Carone, B. R., Fauquier, L., Habib, N., Shea, J. M., Hart, C. E., Li, R., Bock, C., Li, C., Gu, H., Zamore, P. D., Meissner, A., Weng, Z., Hofmann, H. A., Friedman, N., and Rando, O. J. (2010). Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell*, 143(7):1084–1096.

Cech, T. R. and Steitz, J. A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, 157(1):77–94.

Chatzi, C., Zhang, Y., Shen, R., Westbrook, G. L., and Goodman, R. H. (2016). Transcriptional profiling of newly generated dentate granule cells using TU tagging reveals pattern shifts in gene expression during circuit integration. *eNeuro*, 3(1).

Chazaud, C. and Yamanaka, Y. (2016). Lineage specification in the mouse preimplantation embryo. *Development*, 143(7):1063–1074.

Chen, Q., Yan, M., Cao, Z., Li, X., Zhang, Y., Shi, J., Feng, G.-H., Peng, H., Zhang, X., Zhang, Y., Qian, J., Duan, E., Zhai, Q., and Zhou, Q. (2016). Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. *Science*, 351(6271):397–400.

Chen, X., Gu, X., and Zhang, H. (2018). Sidt2 regulates hepatocellular lipid metabolism through autophagy. *J. Lipid Res.*, 59(3):404–415.

Chen, Z. and Zhang, Y. (2019). Loss of DUX causes minor defects in zygotic genome activation and is compatible with mouse development. *Nat. Genet.*, 51(6):947–951.

Chomczynski, P. and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.*, 162(1):156–159.

Chu, C., Zhang, Q. C., da Rocha, S. T., Flynn, R. A., Bharadwaj, M., Calabrese, J. M., Magnuson, T., Heard, E., and Chang, H. Y. (2015). Systematic discovery of Xist RNA binding proteins. *Cell*, 161(2):404–416.

Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R., and Boothroyd, J. C. (2005). Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat. Biotechnol.*, 23(2):232–237.

Clegg, K. B. and Pikó, L. (1983a). Poly(A) length, cytoplasmic adenylation and synthesis of poly(A)+ RNA in early mouse embryos. *Dev. Biol.*, 95(2):331–341.

Clegg, K. B. and Pikó, L. (1983b). Quantitative aspects of RNA synthesis and polyadenylation in 1-cell and 2-cell mouse embryos. *J. Embryol. Exp. Morphol.*, 74:169–182.

Conine, C. C., Sun, F., Song, L., Rivera-Pérez, J. A., and Rando, O. J. (2018). Small RNAs gained during epididymal transit of sperm are essential for embryonic development in mice. *Dev. Cell*, 46(4):470–480.e3.

Crick, F. H., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, 192:1227–1232.

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Dev. Biol.*, 278(2):274–288.

Darwin, C. (2010). *The Variation of Animals and Plants under Domestication*, volume 2 of *Cambridge Library Collection - Darwin, Evolution and Genetics*. Cambridge University Press.

Das, P. P., Bagijn, M. P., Goldstein, L. D., Woolford, J. R., Lehrbach, N. J., Sapetschnig, A., Buhecha, H. R., Gilchrist, M. J., Howe, K. L., Stark, R., Matthews, N., Berezikov, E., Ketting, R. F., Tavaré, S., and Miska, E. A. (2008). Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress tc3 transposon mobility in the caenorhabditis elegans germline. *Mol. Cell*, 31(1):79–90.

De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., Funaya, C., Antony, C., Moreira, P. N., Enright, A. J., and O'Carroll, D. (2011). The endonuclease activity of mili fuels piRNA amplification that silences LINE1 elements. *Nature*, 480(7376):259–263.

De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., and Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.*, 49(6):941–945.

De Iaco, A., Verp, S., Offner, S., and Trono, D. (2019). DUX is a non-essential synchronizer of zygotic genome activation. *bioRxiv*.

de Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., Åström, G., Babina, M., Bertin, N., Burroughs, A. M., Carlisle, A. J., Daub, C. O., Detmar, M., Deviatiiarov, R., Fort, A., Gebhard, C., Goldowitz, D., Guhl, S., Ha, T. J., Harshbarger, J., Hasegawa, A., Hashimoto, K., Herlyn, M., Heutink, P., Hitchens, K. J., Hon, C. C., Huang, E., Ishizu, Y., Kai, C., Kasukawa, T., Klinken, P., Lassmann, T., Lecellier, C.-H., Lee, W., Lizio, M., Makeev, V., Mathelier, A., Medvedeva, Y. A., Mejhert, N., Mungall, C. J., Noma, S., Ohshima, M., Okada-Hatakeyama, M., Persson, H., Rizzu, P., Roudnicky, F., Sætrom, P., Sato, H., Severin, J., Shin, J. W., Swoboda, R. K., Tarui, H., Toyoda, H., Vitting-Seerup, K., Winteringham, L., Yamaguchi, Y., Yasuzawa, K., Yoneda, M., Yumoto, N., Zabierowski, S., Zhang, P. G., Wells, C. A., Summers, K. M., Kawaji, H., Sandelin, A., Rehli, M., FANTOM Consortium, Hayashizaki, Y., Carninci, P., Forrest, A. R. R., and de Hoon, M. J. L. (2017). An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.*, 35(9):872–878.

Deng, Z.-B., Zhuang, X., Ju, S., Xiang, X., Mu, J., Liu, Y., Jiang, H., Zhang, L., Mobley, J., McClain, C., Feng, W., Grizzle, W., Yan, J., Miller, D., Kronenberg, M., and Zhang, H.-G. (2013). Exosome-like nanoparticles from intestinal mucosal cells carry prostaglandin E2 and suppress activation of liver NKT cells. *J. Immunol.*, 190(7):3579–3589.

Doe, B., Brown, E., and Boroviak, K. (2018). Generating CRISPR/Cas9-Derived mutant mice by zygote cytoplasmic injection using an automatic microinjector. *Methods Protoc*, 1(1).

Dölken, L., Ruzsics, Z., Rädle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P., and Koszinowski, U. H. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, 14(9):1959–1972.

Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, 15:215.

Duffy, E. E., Rutenberg-Schoenberg, M., Stark, C. D., Kitchen, R. R., Gerstein, M. B., and Simon, M. D. (2015). Tracking distinct RNA populations using efficient and reversible covalent chemistry. *Mol. Cell*, 59(5):858–866.

Dunoyer, P., Himber, C., Ruiz-Ferrer, V., Alioua, A., and Voinnet, O. (2007). Intra- and intercellular RNA interference in arabidopsis thaliana requires components of the microRNA and heterochromatic silencing pathways. *Nat. Genet.*, 39(7):848–856.

Eguchi, J., Wang, X., Yu, S., Kershaw, E. E., Chiu, P. C., Dushay, J., Estall, J. L., Klein, U., Maratos-Flier, E., and Rosen, E. D. (2011). Transcriptional control of adipose lipid handling by IRF4. *Cell Metab.*, 13(3):249–259.

El-Hefnawy, T., Raja, S., Kelly, L., Bigbee, W. L., Kirkwood, J. M., Luketich, J. D., and Godfrey, T. E. (2004). Characterization of amplifiable, circulating RNA in plasma and its potential as a tool for cancer diagnostics. *Clin. Chem.*, 50(3):564–573.

Ellem, K. A. and Gwatkin, R. B. (1968). Patterns of nucleic acid synthesis in the early mouse embryo. *Dev. Biol.*, 18(4):311–330.

Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., Zhuang, Z., Goldstein, S. R., Weiss, R. A., and Liotta, L. A. (1996). Laser capture microdissection. *Science*, 274(5289):998–1001.

Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E. S., Plath, K., and Guttman, M. (2013). The Xist lncRNA exploits Three-Dimensional genome architecture to spread across the X chromosome. *Science*, 341(August):1–8.

Erhard, F., Baptista, M. A. P., Krammer, T., Hennig, T., Lange, M., Arampatzi, P., Jürges, C. S., Theis, F. J., Saliba, A.-E., and Dölken, L. (2019). scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature*, 571(7765):419–423.

Erickson, T. and Nicolson, T. (2015). Identification of sensory hair-cell transcripts by thiouracil-tagging in zebrafish. *BMC Genomics*, 16:842.

Evsikov, A. V., de Vries, W. N., Peaston, A. E., Radford, E. E., Fancher, K. S., Chen, F. H., Blake, J. A., Bult, C. J., Latham, K. E., Solter, D., and Knowles, B. B. (2004). Systems biology of the 2-cell mouse embryo. *Cytogenet. Genome Res.*, 105(2-4):240–250.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Res.*, 8(3):186–194.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Res.*, 8(3):175–185.

Fadloun, A., Le Gras, S., Jost, B., Ziegler-Birling, C., Takahashi, H., Gorab, E., Carninci, P., and Torres-Padilla, M.-E. (2013). Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. *Nat. Struct. Mol. Biol.*, 20(3):332–338.

Feil, R., Wagner, J., Metzger, D., and Chambon, P. (1997). Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. *Biochem. Biophys. Res. Commun.*, 237(3):752–757.

Feinberg, E. H. and Hunter, C. P. (2003). Transport of dsRNA into cells by the transmembrane protein SID-1. *Science*, 301(5639):1545–1547.

Feng, H., Conneely, K. N., and Wu, H. (2014). A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.*, 42(8):e69.

Fire, A., Albertson, D., Harrison, S. W., and Moerman, D. G. (1991). Production of antisense RNA leads to effective and specific inhibition of gene expression in C. elegans muscle. *Development*, 113(2):503–514.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811.

Flemr, M., Malik, R., Franke, V., Nejepinska, J., Sedlacek, R., Vlahovicek, K., and Svoboda, P. (2013). A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell*, 155(4):807–816.

Frenette, G., Lessard, C., and Sullivan, R. (2002). Selected proteins of "prostasome-like particles" from epididymal cauda fluid are transferred to epididymal caput spermatozoa in bull. *Biol. Reprod.*, 67(1):308–313.

Fu, G. K., Hu, J., Wang, P.-H., and Fodor, S. P. A. (2011). Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl. Acad. Sci. U. S. A.*, 108(22):9026–9031.

Fuda, N. J., Ardehali, M. B., and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–192.

Gao, J., Gu, X., Mahuran, D. J., Wang, Z., and Zhang, H. (2013). Impaired glucose tolerance in a mouse model of Sidt2 deficiency. *PLoS One*, 8(6):e66139.

Gao, J., Zhang, Y., Yu, C., Tan, F., and Wang, L. (2016). Spontaneous nonalcoholic fatty liver disease and ER stress in Sidt2 deficiency mice. *Biochem. Biophys. Res. Commun.*, 476(4):326–332.

Gapp, K., Jawaid, A., Sarkies, P., Bohacek, J., Pelczar, P., Prados, J., Farinelli, L., Miska, E., and Mansuy, I. M. (2014). Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. *Nat. Neurosci.*, 17(5):667–669.

Gay, L., Miller, M. R., Ventura, P. B., Devasthali, V., Vue, Z., Thompson, H. L., Temple, S., Zong, H., Cleary, M. D., Stankunas, K., and Doe, C. Q. (2013). Mouse TU tagging: a chemical/genetic intersectional method for purifying cell type-specific nascent RNA. *Genes Dev.*, 27(1):98–115.

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., Enright, A. J., and Schier, A. F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, 312(5770):75–79.

Gorman, J. and Halvorson, H. (1959). Relationship between protein and nucleic acid synthesis in pseudomonas azotogensis grown in hexetidine. *Arch. Biochem. Biophys.*, 84:462–470.

Goudarzi, M., Berg, K., Pieper, L. M., and Schier, A. F. (2019). Individual long non-coding RNAs have no overt functions in zebrafish embryogenesis, viability and fertility. *Elife*, 8.

Hamatani, T., Carter, M. G., Sharov, A. A., and Ko, M. S. H. (2004). Dynamics of global gene expression changes during mouse preimplantation development. *Dev. Cell*, 6(1):117–131.

Harvey, S. A., Sealy, I., Kettleborough, R., Fenyes, F., White, R., Stemple, D., and Smith, J. C. (2013). Identification of the zebrafish maternal and paternal transcriptomes. *Development*, 140(13):2703–2710.

Heffner, C. S., Herbert Pratt, C., Babiuk, R. P., Sharma, Y., Rockwood, S. F., Donahue, L. R., Eppig, J. T., and Murray, S. A. (2012). Supporting conditional mouse mutagenesis with a comprehensive cre characterization resource. *Nat. Commun.*, 3:1218.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589.

Hendrickson, P. G., Doráis, J. A., Grow, E. J., Whiddon, J. L., Lim, J.-W., Wike, C. L., Weaver, B. D., Pflueger, C., Emery, B. R., Wilcox, A. L., Nix, D. A., Peterson, C. M., Tapscott, S. J., Carrell, D. T., and Cairns, B. R. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.*, 49(6):925–934.

Hershey, A. D., Dixon, J., and Chase, M. (1953). Nucleic acid economy in bacteria infected with bacteriophage t2. i. purine and pyrimidine composition. *J. Gen. Physiol.*, 36(6):777–789.

Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., Wlotzka, W., von Haeseler, A., Zuber, J., and Ameres, S. L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods*, 14(12):1198–1204.

Heyn, P., Kircher, M., Dahl, A., Kelso, J., Tomancak, P., Kalinka, A. T., and Neugebauer, K. M. (2014). The earliest transcribed zygotic genes are short, newly evolved, and different across species. *Cell Rep.*, 6(2):285–292.

Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I., and Zamecnik, P. C. (1958). A soluble ribonucleic acid intermediate in protein synthesis. *J. Biol. Chem.*, 231(1):241–257.

Hug, H. and Schuler, R. (2003). Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J. Theor. Biol.*, 221(4):615–624.

Hupe, M., Li, M. X., Gertow Gillner, K., Adams, R. H., and Stenman, J. M. (2014). Evaluation of TRAP-sequencing technology with a versatile conditional mouse model. *Nucleic Acids Res.*, 42(2):e14.

Hwang, H.-W., Saito, Y., Park, C. Y., Blachère, N. E., Tajima, Y., Fak, J. J., Zucker-Scharff, I., and Darnell, R. B. (2017). cTag-PAPERCLIP reveals alternative polyadenylation promotes Cell-Type specific protein diversity and shifts araf isoforms with microglia activation. *Neuron*, 95(6):1334–1349.e5.

Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646):550–554.

Indra, A. K., Warot, X., Brocard, J., Bornert, J. M., Xiao, J. H., Chambon, P., and Metzger, D. (1999). Temporally-controlled site-specific mutagenesis in the basal layer of the epidermis: comparison of the recombinase activity of the tamoxifen-inducible Cre-ER(T) and Cre-ER(T2) recombinases. *Nucleic Acids Res.*, 27(22):4324–4327.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3(3):318–356.

Jalkanen, J., Kotimäki, M., Huhtaniemi, I., and Poutanen, M. (2006). Novel epididymal protease inhibitors with kazal or WAP family domain. *Biochem. Biophys. Res. Commun.*, 349(1):245–254.

Jao, C. Y. and Salic, A. (2008). Exploring RNA transcription and turnover in vivo by using click chemistry. *Proc. Natl. Acad. Sci. U. S. A.*, 105(41):15779–15784.

Javillier, M. and Fabrykant, M. (1931). Recherches experimentales sur le phosphore sanguin et particulierment sur variations de la phosphatemie. *Bull Soc Chim Biol*, 13:1253–1256.

Johnston, D. S., Jelinsky, S. A., Bang, H. J., DiCandeloro, P., Wilson, E., Kopf, G. S., and Turner, T. T. (2005). The mouse epididymal transcriptome: transcriptional profiling of segmental gene expression in the epididymis. *Biol. Reprod.*, 73(3):404–413.

Julius, M. H., Masuda, T., and Herzenberg, L. A. (1972). Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. *Proc. Natl. Acad. Sci. U. S. A.*, 69(7):1934–1938.

Kageyama, S.-I., Nagata, M., and Aoki, F. (2004). Isolation of nascent messenger RNA from mouse preimplantation embryos. *Biol. Reprod.*, 71(6):1948–1955.

Kamm, R. C. and Smith, A. G. (1972). Nucleic acid concentrations in normal human plasma. *Clin. Chem.*, 18(6):519–522.

Kang, D.-C., Gopalkrishnan, R. V., Wu, Q., Jankowsky, E., Pyle, A. M., and Fisher, P. B. (2002). mda-5: An interferon-inducible putative RNA helicase with double-stranded RNA-dependent ATPase activity and melanoma growth-suppressive properties. *Proc. Natl. Acad. Sci. U. S. A.*, 99(2):637–642.

Kenzelmann, M., Maertens, S., Hergenhahn, M., Kueffer, S., Hotz-Wagenblatt, A., Li, L., Wang, S., Ittrich, C., Lemberger, T., Arribas, R., Jonnakuty, S., Hollstein, M. C., Schmid, W., Gretz, N., Gröne, H. J., and Schütz, G. (2007). Microarray analysis of newly synthesized RNA in cells and animals. *Proc. Natl. Acad. Sci. U. S. A.*, 104(15):6164–6169.

Kigami, D., Minami, N., Takayama, H., and Imai, H. (2003). MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biol. Reprod.*, 68(2):651–654.

Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, 10(2):126–139.

Kisanuki, Y. Y., Hammer, R. E., Miyazaki, J., Williams, S. C., Richardson, J. A., and Yanagisawa, M. (2001). Tie2-Cre transgenic mice: a new model for endothelial cell-lineage analysis in vivo. *Dev. Biol.*, 230(2):230–242.

Kopreski, M. S., Benko, F. A., Kwak, L. W., and Gocke, C. D. (1999). Detection of tumor messenger RNA in the serum of patients with malignant melanoma. *Clin. Cancer Res.*, 5(8):1961–1965.

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T. W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., Lin, H., Matsuda, Y., and Nakano, T. (2004). Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development*, 131(4):839–849.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*.

Lakoski, K. A., Carron, C. P., Cabot, C. L., and Saling, P. M. (1988). Epididymal maturation and the acrosome reaction in mouse sperm: response to zona pellucida develops coincident with modification of M42 antigen. *Biol. Reprod.*, 38(1):221–233.

Lamarck, J. B. (1809). *Zoological Philosophy*. Macmillan.

Lee, H.-C., Gu, W., Shirayama, M., Youngman, E., Conte, Jr, D., and Mello, C. C. (2012). C. elegans piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell*, 150(1):78–87.

Lee, M. T., Bonneau, A. R., Takacs, C. M., Bazzini, A. A., DiVito, K. R., Fleming, E. S., and Giraldez, A. J. (2013). Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature*, 503(7476):360–364.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854.

Lee, Y. S., Shibata, Y., Malhotra, A., and Dutta, A. (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.*, 23(22):2639–2649.

Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., and Steitz, J. A. (1980). Are snRNPs involved in splicing? *Nature*, 283(5743):220–224.

Levin, H. L. and Moran, J. V. (2011). Dynamic interactions between transposable elements and their hosts. *Nat. Rev. Genet.*, 12(9):615–627.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20.

Li, J., Huang, S., Chen, J., Yang, Z., Fei, X., Zheng, M., Ji, C., Xie, Y., and Mao, Y. (2007). Identification and characterization of human uracil phosphoribosyltransferase (UPRTase). *J. Hum. Genet.*, 52(5):415–422.

Li, L., Liu, B., Wapinski, O. L., Tsai, M.-C., Qu, K., Zhang, J., Carlson, J. C., Lin, M., Fang, F., Gupta, R. A., Helms, J. A., and Chang, H. Y. (2013). Targeted disruption of Hotair leads to homeotic transformation and gene derepression. *Cell Rep.*, 5(1):3–12.

Liang, H.-L., Nien, C.-Y., Liu, H.-Y., Metzstein, M. M., Kirov, N., and Rushlow, C. (2008). The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, 456(7220):400–403.

Lim, J., Ha, M., Chang, H., Kwon, S. C., Simanshu, D. K., Patel, D. J., and Kim, V. N. (2014). Uridylation by TUT4 and TUT7 marks mRNA for degradation. *Cell*, 159(6):1365–1376.

Lin, H. and Spradling, a. C. (1997). A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. *Development*, 124(12):2463–2476.

Littlefield, J. W., Keller, E. B., Gross, J., and Zamecnik, P. C. (1955). Studies on cytoplasmic ribonucleoprotein particles from the liver of the rat. *J. Biol. Chem.*, 217(1):111–123.

Lo, K. W., Lo, Y. M., Leung, S. F., Tsang, Y. S., Chan, L. Y., Johnson, P. J., Hjelm, N. M., Lee, J. C., and Huang, D. P. (1999). Analysis of cell-free Epstein-Barr virus associated RNA in the plasma of patients with nasopharyngeal carcinoma. *Clin. Chem.*, 45(8 Pt 1):1292–1294.

Lund, E., Liu, M., Hartley, R. S., Sheets, M. D., and Dahlberg, J. E. (2009). Deadenylation of maternal mRNAs mediated by mir-427 in Xenopus laevis embryos. *RNA*, 15(12):2351–2363.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(Nov):2579–2605.

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, 487(7405):57–63.

Madison, B. B., Dunbar, L., Qiao, X. T., Braunstein, K., Braunstein, E., and Gumucio, D. L. (2002). Cis elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J. Biol. Chem.*, 277(36):33275–33283.

Maria Angelica, C., Carlos, B.-R., Jana, F., Gabriel, L.-B., Anil, K. S., and George, A. C. (2011). MicroRNAs in body fluids—the mix of hormones and biomarkers. *Nat. Rev. Clin. Oncol.*, 8(8):467–477.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12.

Matsushima, W., Herzog, V. A., Neumann, T., Gapp, K., Zuber, J., Ameres, S. L., and Miska, E. A. (2018). SLAM-ITseq: sequencing cell type-specific transcriptomes without cell sorting. *Development*, 145(13).

Matsushima, W., Herzog, V. A., Neumann, T., Gapp, K., Zuber, J., Ameres, S. L., and Miska, E. A. (2019). Sequencing cell-type-specific transcriptomes with SLAM-ITseq. *Nat. Protoc.*, 14(8):2261–2278.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, 45(D1):D183–D189.

Mintz, B. (1964). SYNTHETIC PROCESSES AND EARLY DEVELOPMENT IN THE MAMMALIAN EGG. *J. Exp. Zool.*, 157:85–100.

Mishima, Y. and Tomari, Y. (2016). Codon usage and 3′ UTR length determine maternal mRNA stability in zebrafish. *Mol. Cell*, 61(6):874–885.

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B., and Tewari, M. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl. Acad. Sci. U. S. A.*, 105(30):10513–10518.

Molnar, A., Melnyk, C. W., Bassett, A., Hardcastle, T. J., Dunn, R., and Baulcombe, D. C. (2010). Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. *Science*, 328(5980):872–875.

Morgan, M., Much, C., DiGiacomo, M., Azzi, C., Ivanova, I., Vitsios, D. M., Pistolic, J., Collier, P., Moreira, P. N., Benes, V., Enright, A. J., and O'Carroll, D. (2017). mRNA 3′ uridylation and poly(A) tail length sculpt the mammalian maternal transcriptome. *Nature*, 548(7667):347–351.

Muhar, M., Ebert, A., Neumann, T., Umkehrer, C., Jude, J., Wieshofer, C., Rescheneder, P., Lipp, J. J., Herzog, V. A., Reichholf, B., Cisneros, D. A., Hoffmann, T., Schlapansky, M. F., Bhat, P., von Haeseler, A., Köcher, T., Obenauf, A. C., Popow, J., Ameres, S. L., and Zuber, J. (2018). SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science*, 360(6390):800–805.

Murai, S., Stein, P., Buffone, M. G., Yamashita, S., and Schultz, R. M. (2010). Recruitment of Orc6l, a dormant maternal mRNA in mouse oocytes, is essential for DNA replication in 1-cell embryos. *Dev. Biol.*, 341(1):205–212.

Murray, R., Rodwell, V., Bender, D., Botham, K., Weil, P., and Kennelly, P. (2009). *Harper's Illustrated Biochemistry, 28th Edition*. Lange medical book. McGraw-Hill Education.

Neumann, T., Herzog, V. A., Muhar, M., von Haeseler, A., Zuber, J., Ameres, S. L., and Rescheneder, P. (2019). Quantification of experimentally induced nucleotide conversions in high-throughput sequencing datasets. *BMC Bioinformatics*, 20(1):258.

Ng, E. K. O., Tsui, N. B. Y., Lam, N. Y. L., Chiu, R. W. K., Yu, S. C. H., Wong, S. C. C., Lo, E. S. F., Rainer, T. H., Johnson, P. J., and Lo, Y. M. D. (2002). Presence of filterable and nonfilterable mRNA in the plasma of cancer patients and healthy individuals. *Clin. Chem.*, 48(8):1212–1217.

Ng, S.-F., Lin, R. C. Y., Laybutt, D. R., Barres, R., Owens, J. a., and Morris, M. J. (2010). Chronic high-fat diet in fathers programs beta-cell dysfunction in female rat offspring. *Nature*, 467(7318):963–966.

Nguyen, T. A., Smith, B. R. C., Elgass, K. D., Creed, S. J., Cheung, S., Tate, M. D., Belz, G. T., Wicks, I. P., Masters, S. L., and Pang, K. C. (2019). SIDT1 localizes to endolysosomes and mediates Double-Stranded RNA transport into the cytoplasm. *J. Immunol.*, 202(12):3483–3492.

Nguyen, T. A., Smith, B. R. C., Tate, M. D., Belz, G. T., Barrios, M. H., Elgass, K. D., Weisman, A. S., Baker, P. J., Preston, S. P., Whitehead, L., Garnham, A., Lundie, R. J., Smyth, G. K., Pellegrini, M., O'Keeffe, M., Wicks, I. P., Masters, S. L., Hunter, C. P., and Pang, K. C. (2017). SIDT2 transports extracellular dsRNA into the cytoplasm for innate immune recognition. *Immunity*, 47(3):498–509.e6.

Noller, H. F., Hoffarth, V., and Zimniak, L. (1992). Unusual resistance of peptidyl transferase to protein extraction procedures. *Science*, 256(5062):1416–1419.

Nuez, I. and Félix, M.-A. (2012). Evolution of susceptibility to ingested double-stranded RNAs in Caenorhabditis nematodes. *PLoS One*, 7(1):e29811.

O'Flanagan, C. H., Campbell, K. R., Zhang, A. W., Kabeer, F., Lim, J. L. P., Biele, J., Eirew, P., Lai, D., McPherson, A., Kong, E., Bates, C., Borkowski, K., Wiens, M., Hopkins, J., Hewitson, B., Ceglia, N., Moore, R., Mungall, A. J., McAlpine, J. N., The CRUK IMAXT Grand Challenge Team, Shah, S. P., and Aparicio, S. (2019). Dissociation of solid tumour tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *bioRxiv*.

Okaty, B. W., Sugino, K., and Nelson, S. B. (2011). A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. *PLoS One*, 6(1):e16493.

Palauqui, J. C., Elmayan, T., De Borne, F. D., Crete, P., Charles, C., and Vaucheret, H. (1996). Frequencies, timing, and spatial patterns of Co-Suppression of nitrate reductase and nitrite reductase in transgenic tobacco plants. *Plant Physiol.*, 112(4):1447–1456.

Palauqui, J. C., Elmayan, T., Pollien, J. M., and Vaucheret, H. (1997). Systemic acquired silencing: transgene-specific post-transcriptional silencing is transmitted by grafting from silenced stocks to non-silenced scions. *EMBO J.*, 16(15):4738–4745.

Pardee, A. B. (2002). PaJaMas in paris. *Trends Genet.*, 18(11):585–587.

Pardee, A. B., Jacob, F., and Monod, J. (1959). The genetic control and cytoplasmic expression of "inducibility" in the synthesis of β-galactosidase by e. coli. *J. Mol. Biol.*, 1(2):165–178.

Pearson, K. (1901). LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Peaston, A. E., Evsikov, A. V., Graber, J. H., de Vries, W. N., Holbrook, A. E., Solter, D., and Knowles, B. B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell*, 7(4):597–606.

Percharde, M., Lin, C.-J., Yin, Y., Guan, J., Peixoto, G. A., Bulut-Karslioglu, A., Biechele, S., Huang, B., Shen, X., and Ramalho-Santos, M. (2018). A LINE1-Nucleolin partnership regulates early development and ESC identity. *Cell*, 174(2):391–405.e19.

Pfefferkorn, E. R. (1978). Toxoplasma gondii: the enzymic defect of a mutant resistant to 5-fluorodeoxyuridine. *Exp. Parasitol.*, 44(1):26–35.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., and Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.*, 8(1):10950.

Pham, T. V., Piersma, S. R., Warmoes, M., and Jimenez, C. R. (2010). On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics*, 26(3):363–369.

Radford, E. J., Ito, M., Shi, H., Corish, J. A., Yamazawa, K., Isganaitis, E., Seisenberger, S., Hore, T. A., Reik, W., Erkek, S., Peters, A. H. F. M., Patti, M.-E., and Ferguson-Smith, A. C. (2014). In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. *Science*, 345(6198):1255903.

Richardson, G. M., Lannigan, J., and Macara, I. G. (2015). Does FACS perturb gene expression? *Cytometry*, 87(2):166–175.

Ridder, K., Keller, S., Dams, M., Rupp, A.-K., Schlaudraff, J., Del Turco, D., Starmann, J., Macas, J., Karpova, D., Devraj, K., Depboylu, C., Landfried, B., Arnold, B., Plate, K. H., Höglinger, G., Sültmann, H., Altevogt, P., and Momma, S. (2014). Extracellular vesicle-mediated transfer of genetic information between the hematopoietic system and the brain in response to inflammation. *PLoS Biol.*, 12(6):e1001874.

Riml, C., Amort, T., Rieder, D., Gasser, C., Lusser, A., and Micura, R. (2017). Osmium-Mediated transformation of 4-thiouridine to cytidine as key to study RNA dynamics by sequencing. *Angew. Chem. Int. Ed Engl.*, 56(43):13479–13483.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., and Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rostovtsev, V. V., Green, L. G., Fokin, V. V., and Sharpless, K. B. (2002). A stepwise huisgen cycloaddition process: copper (I)-catalyzed regioselective "ligation" of azides and terminal alkynes. *Angew. Chem. Int. Ed.*, 41(14):2596–2599.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., and Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17:125.

Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C., and Simon, M. D. (2018). TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods*, 15(3):221–225.

Sedlazeck, F. J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, 29(21):2790–2791.

Selleri, L., Bartolomei, M. S., Bickmore, W. A., He, L., Stubbs, L., Reik, W., and Barsh, G. S. (2016). A Hox-Embedded long noncoding RNA: Is it all hot air? *PLoS Genet.*, 12(12):e1006485.

Sharma, U., Conine, C. C., Shea, J. M., Boskovic, A., Derr, A. G., Bing, X. Y., Belleannee, C., Kucukural, A., Serra, R. W., Sun, F., Song, L., Carone, B. R., Ricci, E. P., Li, X. Z., Fauquier, L., Moore, M. J., Sullivan, R., Mello, C. C., Garber, M., and Rando, O. J. (2016). Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science*, 351(6271):391–396.

Sharma, U., Sun, F., Conine, C. C., Reichholf, B., Kukreja, S., Herzog, V. A., Ameres, S. L., and Rando, O. J. (2018). Small RNAs are trafficked from the epididymis to developing mammalian sperm. *Dev. Cell*, 0(0).

Shen, B. and Goodman, H. M. (2004). Uridine addition after microRNA-directed cleavage. *Science*, 306(5698):997.

Simons, R. W. and Kleckner, N. (1983). Translational control of IS10 transposition. *Cell*, 34(2):683–691.

Siomi, M. C., Sato, K., Pezic, D., and Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.*, 12(4):246–258.

Stein, P., Zeng, F., Pan, H., and Schultz, R. M. (2005). Absence of non-specific effects of RNA interference triggered by long double-stranded RNA in mouse oocytes. *Dev. Biol.*, 286(2):464–471.

Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H., and Bartel, D. P. (2014). poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*.

Suh, N., Baehner, L., Moltzahn, F., Melton, C., Shenoy, A., Chen, J., and Blelloch, R. (2010). MicroRNA function is globally suppressed in mouse oocytes and early embryos. *Curr. Biol.*, 20(3):271–277.

Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A., and Li, W. (2014). MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.*, 15(2):R38.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800.

Szvicsek, Z., Oszvald, Á., Szabó, L., Sándor, G. O., Kelemen, A., Soós, A. Á., Pálóczi, K., Harsányi, L., Tölgyes, T., Dede, K., Bursics, A., Buzás, E. I., Zeöld, A., and Wiener, Z. (2019). Extracellular vesicle release from intestinal organoids is modulated by apc mutation and other colorectal cancer progression factors. *Cell. Mol. Life Sci.*, 76(12):2463–2476.

Takahashi, M., Contu, V. R., Kabuta, C., Hase, K., Fujiwara, Y., Wada, K., and Kabuta, T. (2017). SIDT2 mediates gymnosis, the uptake of naked single-stranded oligonucleotides into living cells. *RNA Biol.*, 14(11):1534–1543.

Tamm, I., Hand, R., and Caliguiri, L. A. (1976). Action of dichlorobenzimidazole riboside on RNA synthesis in L-929 and HeLa cells. *J. Cell Biol.*, 69(2):229–240.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, 6(5):377–382.

Thomou, T., Mori, M. A., Dreyfuss, J. M., Konishi, M., Sakaguchi, M., Wolfrum, C., Rao, T. N., Winnay, J. N., Garcia-Martin, R., Grinspoon, S. K., Gorden, P., and Kahn, C. R. (2017). Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature*, 542(7642):450–455.

Timmons, L., Court, D. L., and Fire, A. (2001). Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. *Gene*, 263(1-2):103–112.

Tornøe, C. W., Christensen, C., and Meldal, M. (2002). Peptidotriazoles on solid phase: [1,2,3]-triazoles by regiospecific copper(I)-catalyzed 1,3-dipolar cycloadditions of terminal alkynes to azides. *J. Org. Chem.*, 67(9):3057–3064.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M. J. V., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):516–520.

Uddin, M., Altmann, G. G., and Leblond, C. P. (1984). Radioautographic visualization of differences in the pattern of [3H]uridine and [3H]orotic acid incorporation into the RNA of migrating columnar cells in the rat small intestine. *J. Cell Biol.*, 98(5):1619–1629.

Valadi, H., Ekström, K., Bossios, A., Sjöstrand, M., Lee, J. J., and Lötvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat. Cell Biol.*, 9(6):654–659.

Valdes, V. J., Athie, A., Salinas, L. S., Navarro, R. E., and Vaca, L. (2012). CUP-1 is a novel protein involved in dietary cholesterol uptake in Caenorhabditis elegans. *PLoS One*, 7(3):e33962.

van den Brink, S. C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C. S., Robin, C., and van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods*, 14(10):935–936.

van Dongen, S., Abreu-Goodger, C., and Enright, A. J. (2008). Detecting microRNA binding and siRNA off-target effects from expression data. *Nat. Methods*, 5(12):1023–1025.

Voinnet, O. and Baulcombe, D. C. (1997). Systemic signalling in gene silencing. *Nature*, 389(6651):553.

Voinnet, O., Vain, P., Angell, S., and Baulcombe, D. C. (1998). Systemic spread of sequence-specific transgene RNA degradation in plants is initiated by localized introduction of ectopic promoterless DNA. *Cell*, 95(2):177–187.

Wansink, D. G., Schul, W., van der Kraan, I., van Steensel, B., van Driel, R., and de Jong, L. (1993). Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. *J. Cell Biol.*, 122(2):283–293.

Warner, C. M. and Versteegh, L. R. (1974). In vivo and in vitro effect of α-amanitin on preimplantation mouse embryo RNA polymerase. *Nature*, 248:678.

Weismann, A. (1892). *Essays upon heredity and kindred biological problems*. Oxford.

Whiddon, J. L., Langford, A. T., Wong, C.-J., Zhong, J. W., and Tapscott, S. J. (2017). Conservation and innovation in the DUX4-family gene network. *Nat. Genet.*, 49(6):935–940.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862.

Wilhelm, D. and Koopman, P. (2006). The makings of maleness: towards an integrated view of male sexual development. *Nat. Rev. Genet.*, 7(8):620–631.

Winston, W. M., Molodowitch, C., and Hunter, C. P. (2002). Systemic RNAi in C. elegans requires the putative transmembrane protein SID-1. *Science*, 295(5564):2456–2459.

Winston, W. M., Sutherlin, M., Wright, A. J., Feinberg, E. H., and Hunter, C. P. (2007). Caenorhabditis elegans SID-2 is required for environmental RNA interference. *Proc. Natl. Acad. Sci. U. S. A.*, 104(25):10565–10570.

Wolfrum, C., Shi, S., Jayaprakash, K. N., Jayaraman, M., Wang, G., Pandey, R. K., Rajeev, K. G., Nakayama, T., Charrise, K., Ndungo, E. M., Zimmermann, T., Koteliansky, V., Manoharan, M., and Stoffel, M. (2007). Mechanisms and optimization of in vivo delivery of lipophilic siRNAs. *Nat. Biotechnol.*, 25(10):1149–1157.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., Liu, J.-Y., Horvath, S., and Fan, G. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464):593–597.

Yoneyama, M., Kikuchi, M., Natsukawa, T., Shinobu, N., Imaizumi, T., Miyagishi, M., Taira, K., Akira, S., and Fujita, T. (2004). The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nat. Immunol.*, 5(7):730–737.

Zhang, Y., Chen, K., Sloan, S. A., Bennett, M. L., Scholze, A. R., O'Keeffe, S., Phatnani, H. P., Guarnieri, P., Caneda, C., Ruderisch, N., Deng, S., Liddelow, S. A., Zhang, C., Daneman, R., Maniatis, T., Barres, B. A., and Wu, J. Q. (2014). An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.*, 34(36):11929–11947.

Zhao, J., Sun, B. K., Erwin, J. A., Song, J.-J., and Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, 322(5902):750–756.

Zhou, D., Suzuki, T., Asami, M., and Perry, A. C. F. (2019). Caput epididymidal mouse sperm support full development. *Dev. Cell*, 50(1):5–6.

Zhuang, Y. and Weiner, A. M. (1986). A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell*, 46(6):827–835.

Zook, J. M., Samarov, D., McDaniel, J., Sen, S. K., and Salit, M. (2012). Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. *PLoS One*, 7(7):e41356.

Appendix A

RNA-seq quality check

A.1 RNA-seq on *Tie2-Cre* mice

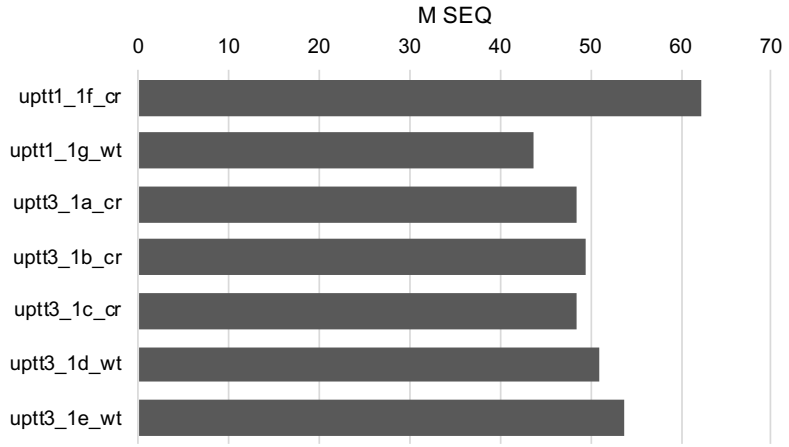


Fig. A.1 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre^- mice and "cr" ID are from Cre^+ mice.

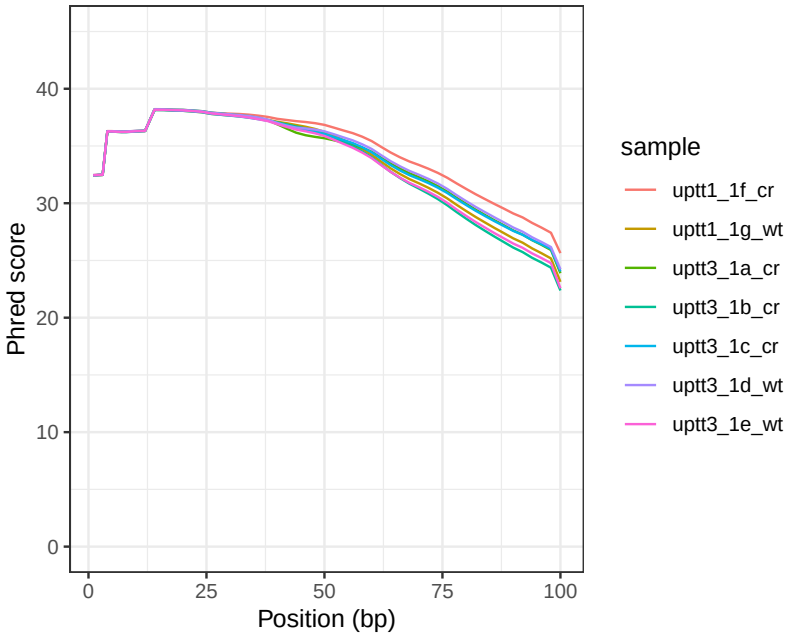


Fig. A.2 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

A.2 RNA-seq on *Vil-Cre* mice

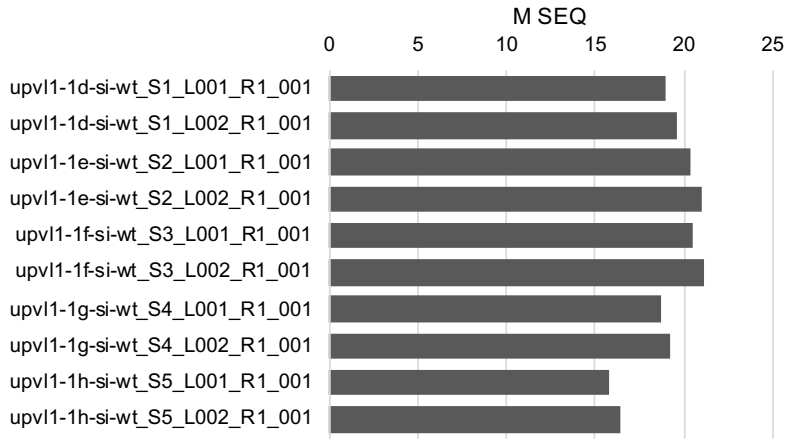


Fig. A.3 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

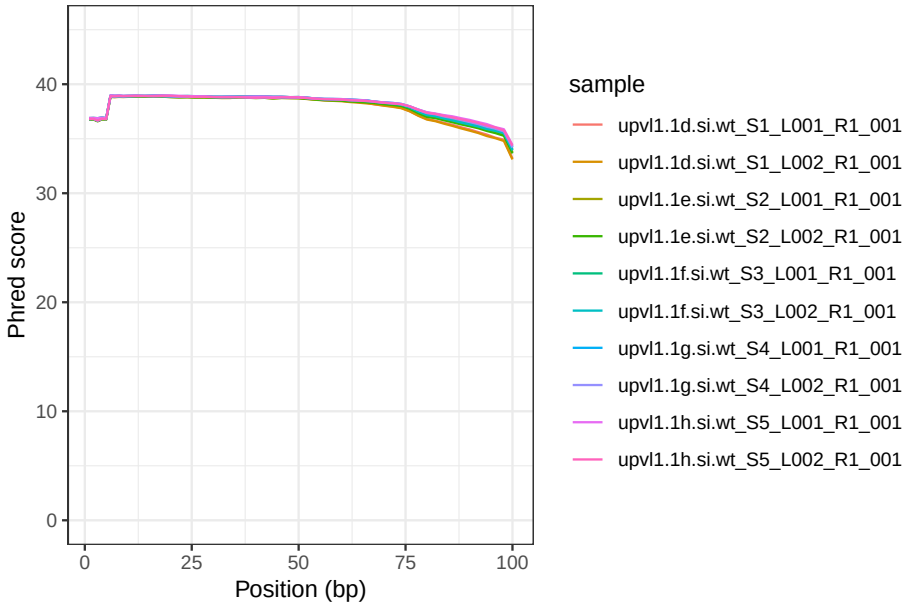


Fig. A.4 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

A.3 RNA-seq on *Adipoq-Cre* mice

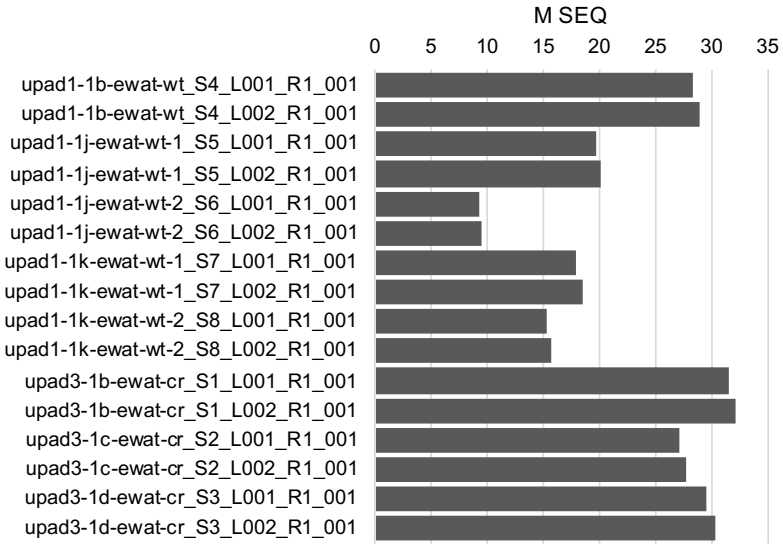


Fig. A.5 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice. Note that the second and the third samples were assigned two barcodes.

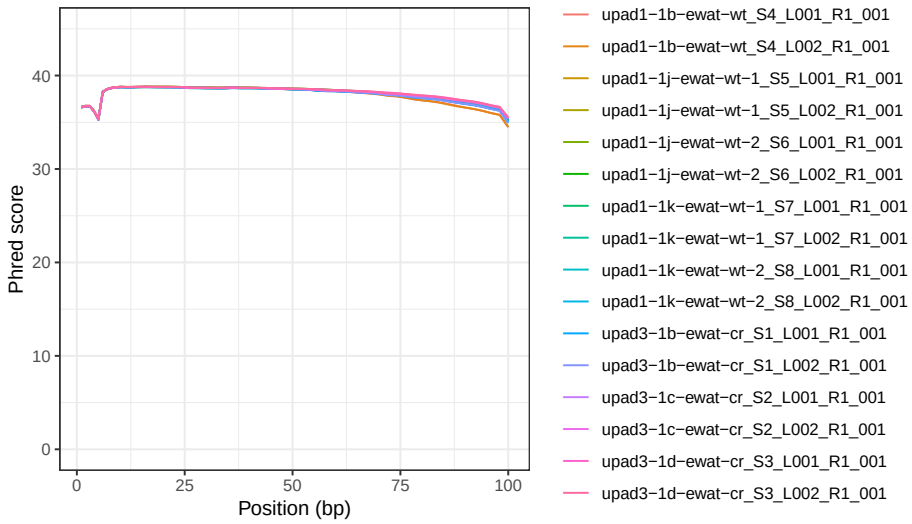


Fig. A.6 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

A.4 Small RNA-seq on WT mice

Fig. A.7 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre$^-$ mice and "cr" ID are from Cre$^+$ mice. Note that the second and the third samples were assigned two barcodes.

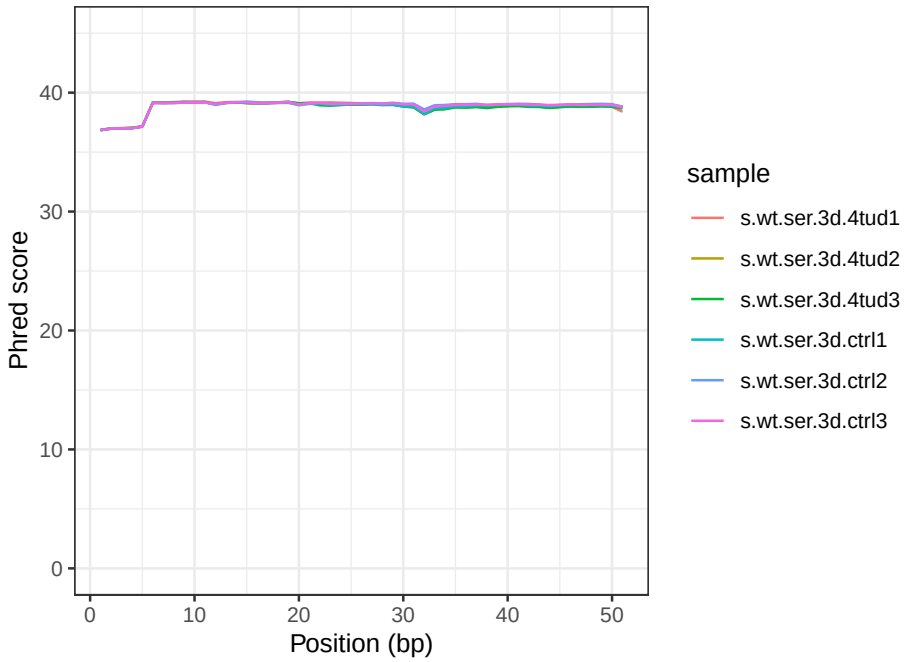


Fig. A.8 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre$^-$ mice and "cr" ID are from Cre$^+$ mice.

A.5 Small RNA-seq on *Vil-Cre* mice

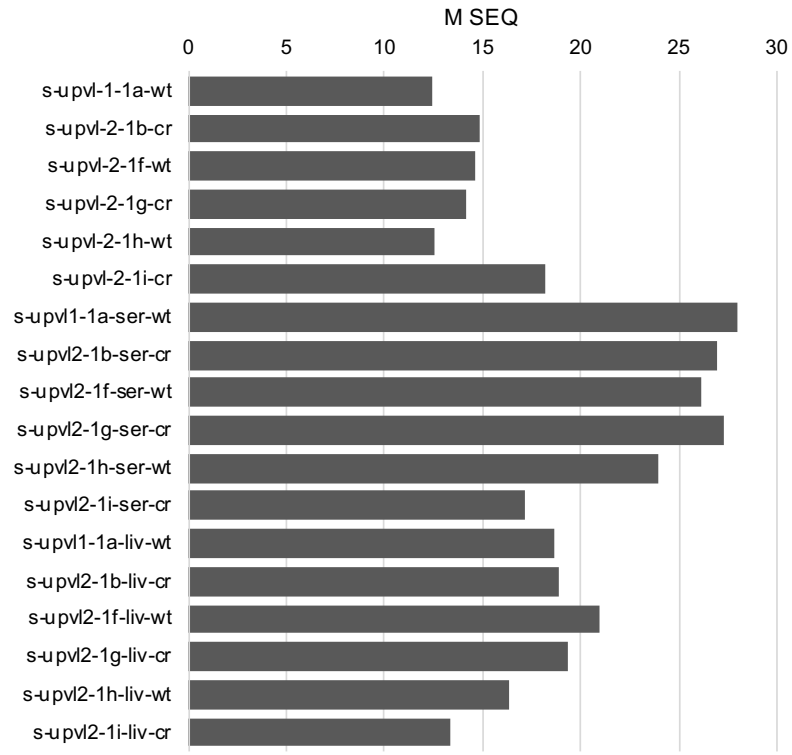


Fig. A.9 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre^- mice and "cr" ID are from Cre^+ mice.

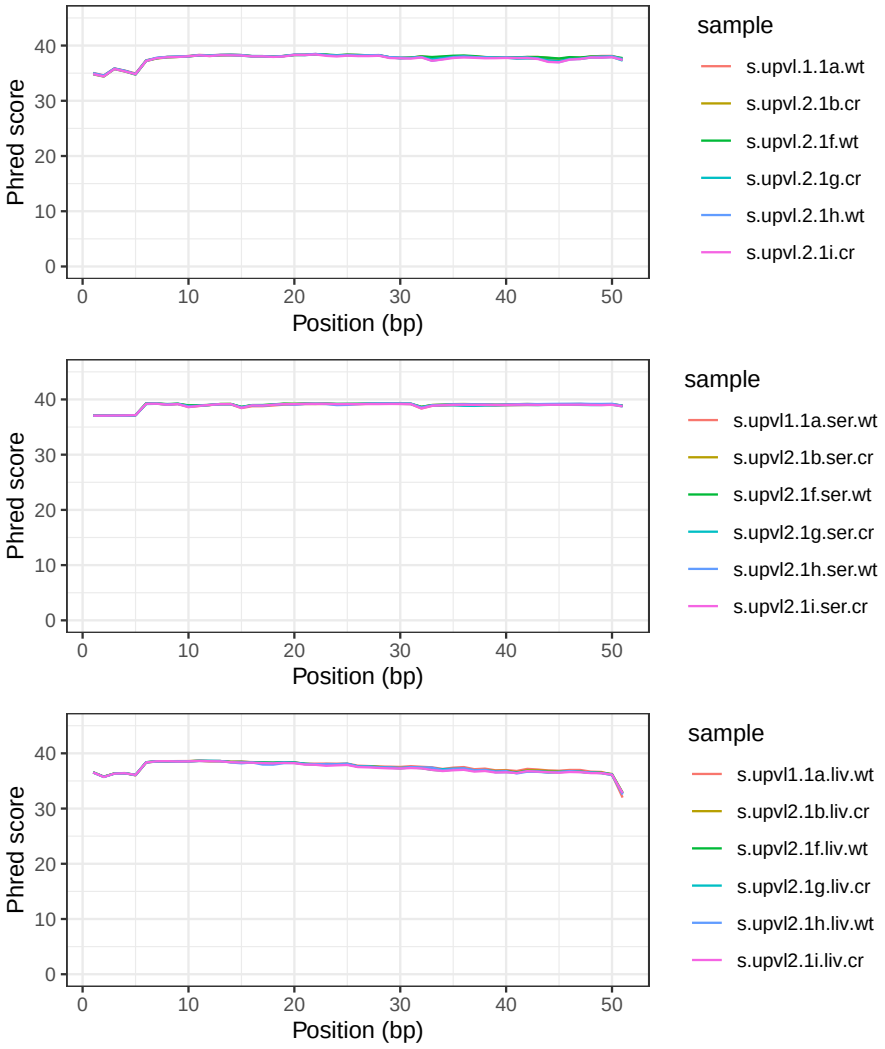


Fig. A.10 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

A.6 Small RNA-seq on *Adipoq-Cre* mice

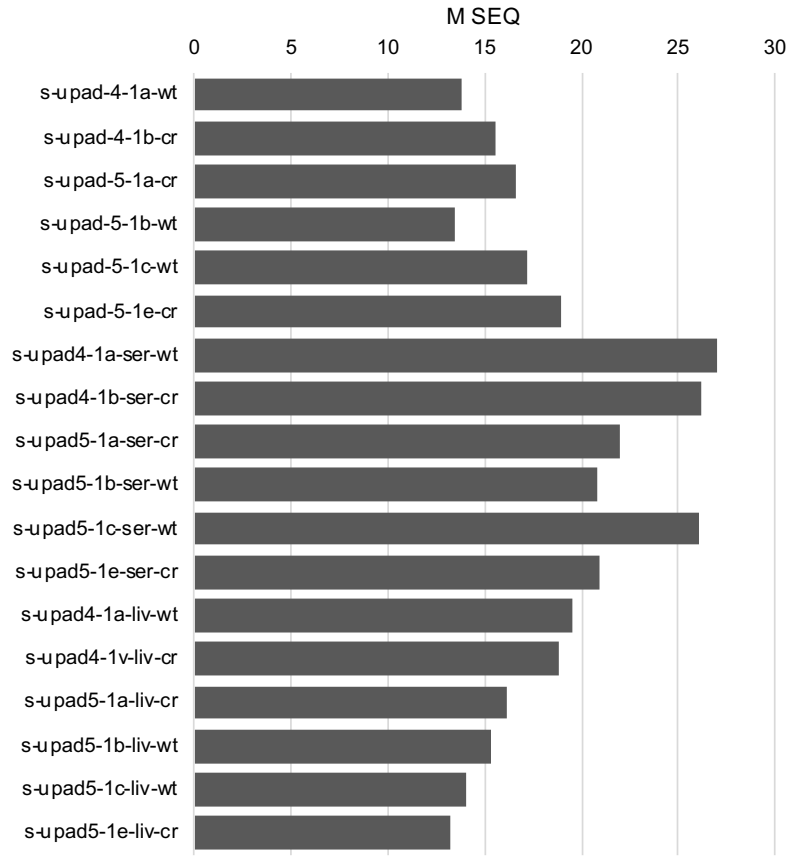


Fig. A.11 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre^- mice and "cr" ID are from Cre^+ mice. Note that the second and the third samples were assigned two barcodes.

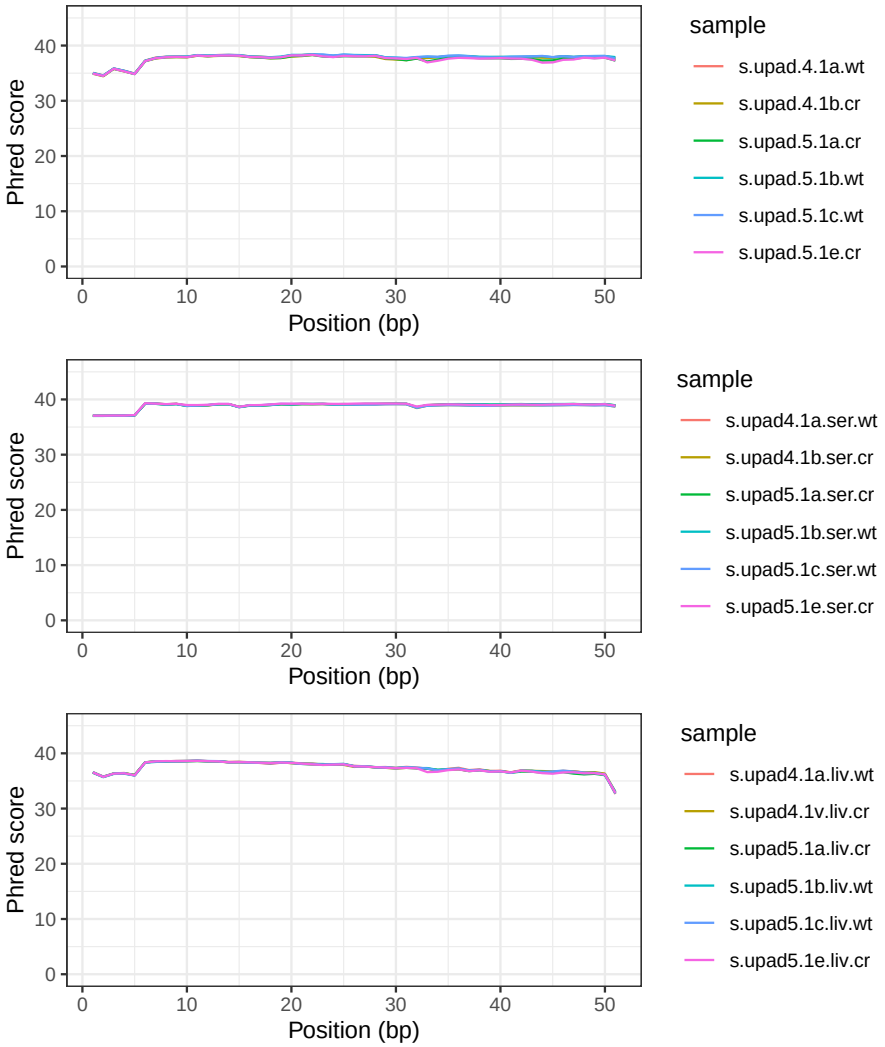


Fig. A.12 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

A.7 RNA-seq on *Spink8-Cre* mice

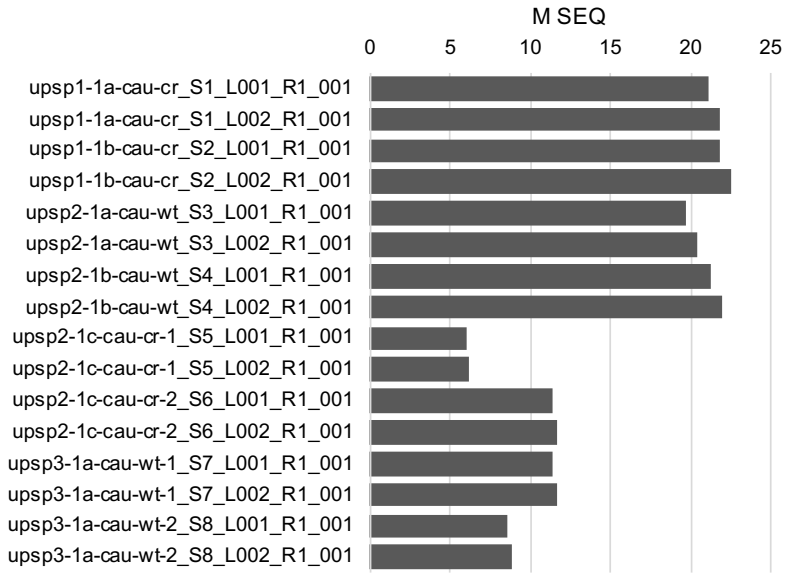


Fig. A.13 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre^- mice and "cr" ID are from Cre^+ mice. Note that the second and the third samples were assigned two barcodes.

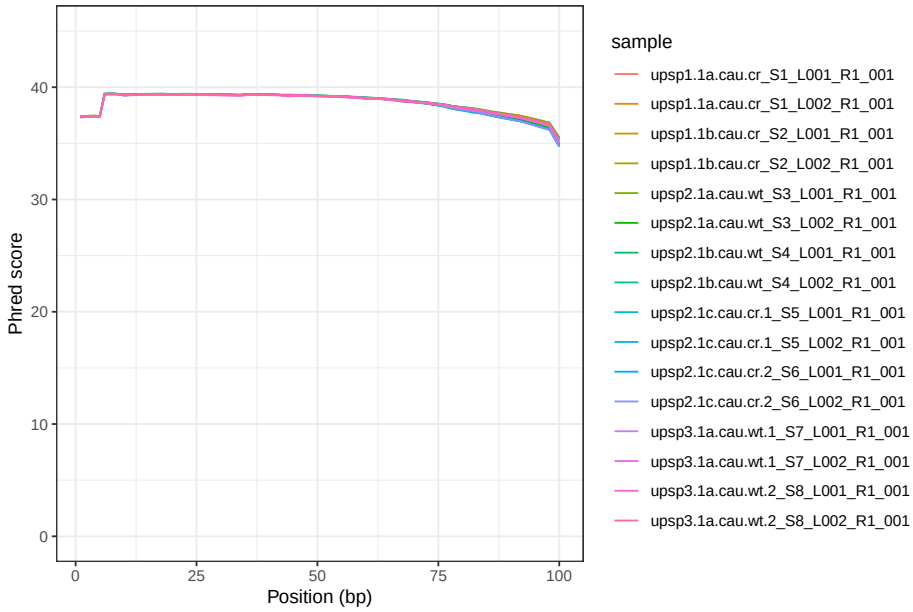


Fig. A.14 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

A.8 Small RNA-seq on *Spink8-Cre* mice

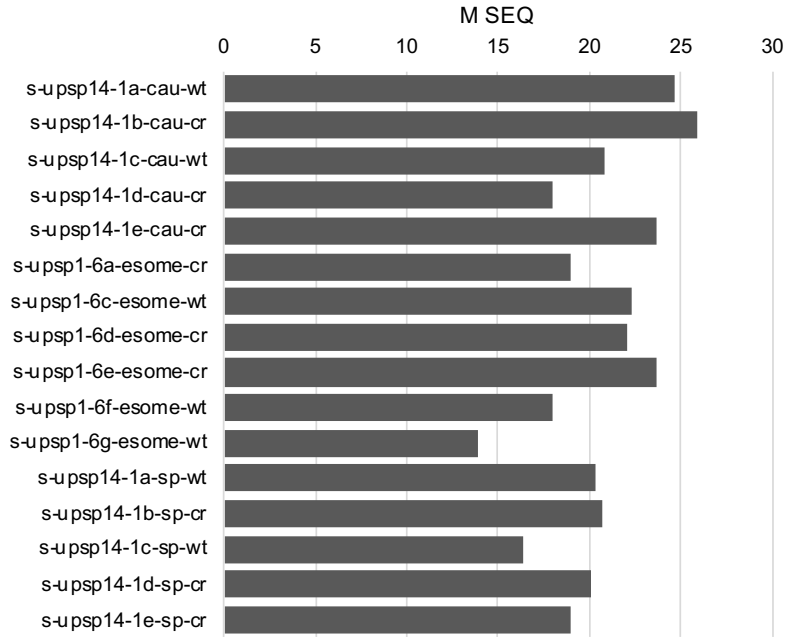


Fig. A.15 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre^- mice and "cr" ID are from Cre^+ mice. Note that the second and the third samples were assigned two barcodes.

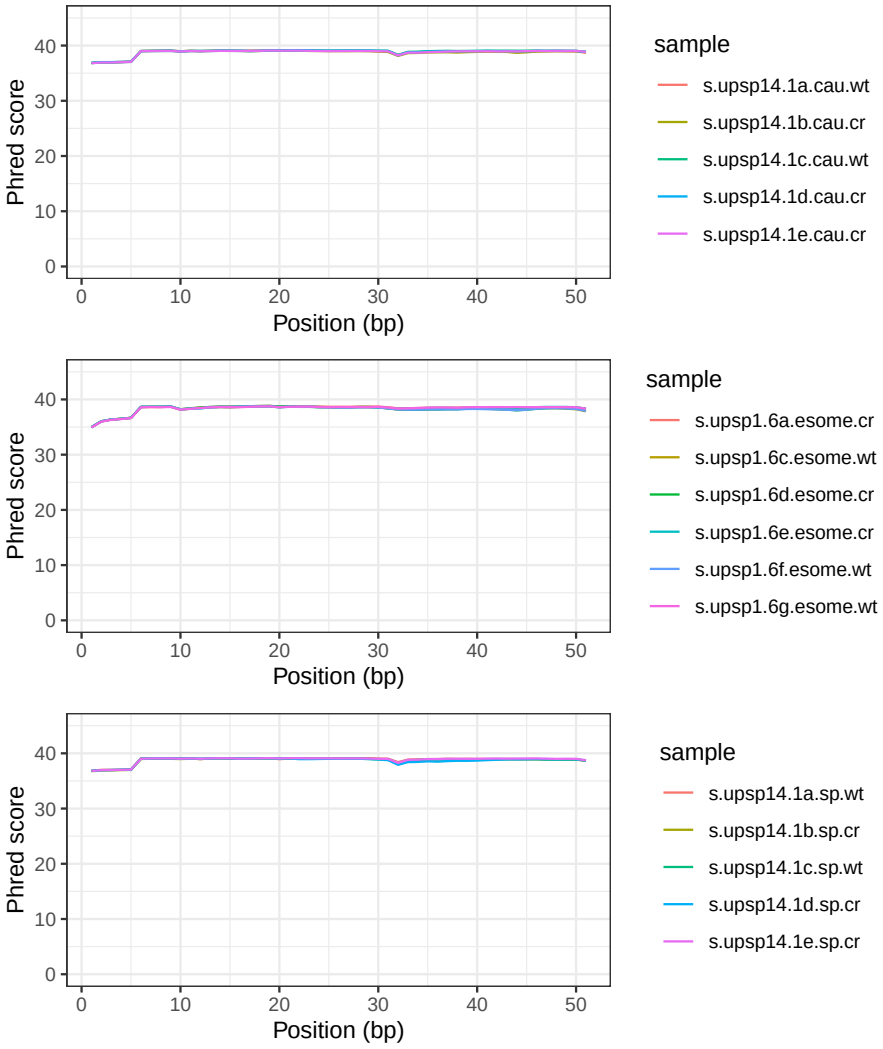


Fig. A.16 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

A.9 RNA-seq on 2C mouse embryo

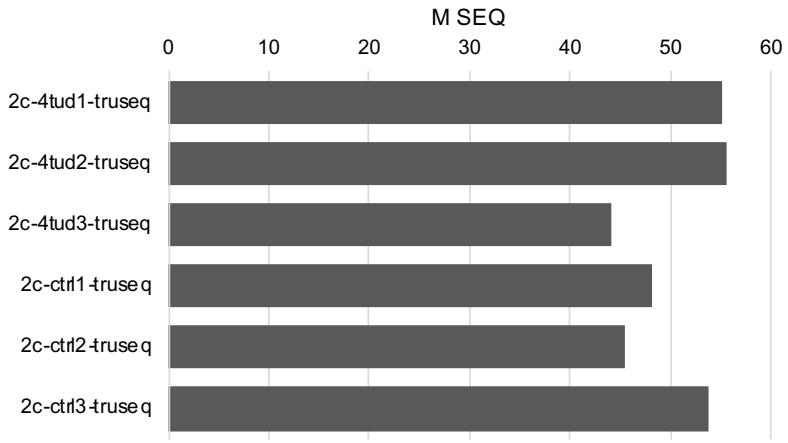


Fig. A.17 Read count obtained for each library
The number of reads obtained in each library is shown. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice. Note that the second and the third samples were assigned two barcodes.

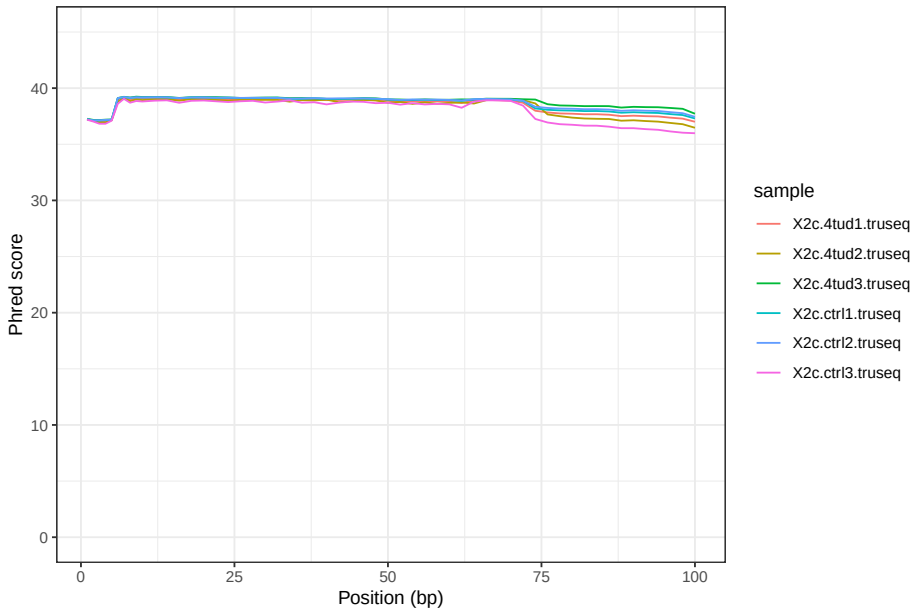


Fig. A.18 Phred score across read for each library
Phred score for each position of read is summarised for each library. Samples with "wt" ID are from Cre^{-} mice and "cr" ID are from Cre^{+} mice.

Appendix B

SLAMseq analysis script

1. Install and load required packages.

```
install.packages(c("dplyr", "tidyr", "ibb"))
library("dplyr")
library("tidyr")
library("ibb")
```

2. Read all the count tables and combine them into one data table.

```
ls <- list.files(pattern = "_tcount.tsv")
df <- do.call(rbind, Map("cbind", lapply(ls, read.delim, skip = 2,
      header = T), sample = gsub("//..*", "", ls)))
```

3. Reshape the data for ibb to perform the beta-binomial test.

```
dfsprd <- df %>%
   select(sample, Name, CoverageOnTs, ConversionsOnTs) %>%
   gather(variable, value, CoverageOnTs, ConversionsOnTs) %>%
   unite(var, variable, sample) %>%
   group_by(var) %>%
   mutate(id = 1:n()) %>%
   spread(var, value)
```

4. Remove genes with no T>C in any of the samples.

```
dfexp <- dfsprd[apply(dfsprd %>%
select(starts_with("CoverageOnTs")), 1, function(z) !any(z == 0)),]
```

5. Specify sample groups as they appear in the "dfexp" columns.

```
dftc <- dfexp %>%
   select(starts_with("ConversionsOnTs"))

dftotal <- dfexp %>%
   select(starts_with("CoverageOnTs"))

group <- c("cr","wt","cr","cr","cr","wt","wt")
```

6. Run `bb.test()` function to perform the beta-binomial test.

```
bbtest <- bb.test(dftc, dftotal, group, n.threads = 0)
```

7. Write a table as a file with *P*-values and FDR (Benjamini-Hochberg's procedure) calculated by beta-binomial test.

```
dfexp$pval <- bbtest$p.value

adj <- dfexp %>%
   mutate(BH = p.adjust(pval, method = "BH"))

write.table(adj, file = "ibb.txt", sep = "/t", quote = F)
```

Appendix C

My publications

Publications to date related to my PhD.

Matsushima, W., Herzog, V. A., Neumann, T., Gapp, K., Zuber, J., Ameres, S. L., and Miska, E. A. (2018). SLAM-ITseq: sequencing cell type-specific transcriptomes without cell sorting. *Development*, 145(13).

Gapp, K., van Steenwyk, G., Germain, PL., **Matsushima, W.**, Rudolph, K. L. M., Manuella, F., Roszkowski, M., Vernaz, G., Ghosh, T., Pelczar, P., Mansuy, I. M., Miska, E. A. (2018). Alterations in sperm long RNA contribute to the epigenetic inheritance of the effects of postnatal trauma. *Mol. Psychiatry*., doi: 10.1038/s41380-018-0271-6

Matsushima, W.*, Brink, K.*, Schroeder, J., Miska, E. A., Gapp, K. (2019). Mature sperm small-RNA profile in the sparrow: Implications for transgenerational effects of age on fitness. *Environ. Epigenet.*, 5 (2): dvz007. *equal contributions

Matsushima, W., Herzog, V. A., Neumann, T., Gapp, K., Zuber, J., Ameres, S. L., and Miska, E. A. (2019). Sequencing cell-type-specific transcriptomes with SLAM-ITseq. *Nat. Protoc.*, 14(8):2261–2278.